Author's response to reviews

Title: Generation of a strand-specific, genomic tiling array of the human Major Histocompatibility Complex (MHC) and its application for DNA methylation analysis.

Authors:

Eleni M Tomazou (emt@sanger.ac.uk)
Vardhman K Rakyan (v.rakyan@qmul.ac.uk)
Gregory Lefebvre (gl2@sanger.ac.uk)
Robert Andrews (rma@sanger.ac.uk)
Peter Ellis (pde@sanger.ac.uk)
David K Jackson (dj3@sanger.ac.uk)
Cordelia Langford (cfl@sanger.ac.uk)
Matthew D Francis (mdf@sanger.ac.uk)
Liselotte Backdahl (l.backdahl@ucl.ac.uk)
Marcos Miretti (mm3@sanger.ac.uk)
Penny Coggill (pcc@sanger.ac.uk)
Diego Ottaviani (Diego.Ottaviani@cancer.org.uk)
Denise Sheer (d.sheer@qmul.ac.uk)
Adele Murrell (Adele.Murrell@cancer.org.uk)
Stephan Beck (s.becl@ucl.ac.uk)

Version: 2 Date: 18 May 2008

Author's response to reviews: see over
May 19, 2008

MS: 8769505961831827

Dear Editor,

Thank you very much for the opportunity to revise our manuscript. Please find below our point-by-point responses to the comments raised by the two reviewers.

We trust that our revised manuscript is now acceptable for publication in *BMC Medical Genomics* and look forward to hearing from you.

Sincerely,

Stephan Beck

**Reviewer’s report 1**

**Comment 1:** The title says that the manuscript describes a strand-specific genomic tiling array, although no data are shown which demonstrate strand-specificity, thus this should be omitted from the title.

**Response 1:** As requested, we have removed ‘strand-specificity’ from the title. However, as the array is stand-specific, we kept a reference to this effect in the text to indicate to potential users that this resource allows for strand-specific analyses to be carried out.

**Comment 2:** p. 13 first paragraph (Figure 2): The authors do not comment on the obvious low methylation observed for the class II region.

**Response 2:** While the MeDIP assay is suitable to detect differences between samples (e.g. the identification of differentially methylated regions as described here), it is not suitable to quantify methylation within a sample. To overcome this current limitation of the MeDIP assay requires further work which is in progress and discussed on page 17. Therefore, it would be premature to comment on the actual methylation level of the class II or any other region. Consequently, we have removed the sentence “In concordance with the ubiquitous expression of HLA class I genes, DNA methylation is lowest around HLA-A, B and C.” from page 13.
**Comment 3:** p. 13, second paragraph and Figure 3a: The spacing (steps of 5?) of the columns is not clear to me. If 9% of the probes are repeat-free (= repeats 0%) the 12-% column would be at 95%, the text says 12% consist of repeat sequences only (=100%).

**Response 3:** We agree and have made the appropriate changes in the text (page 14) and legend to Fig 3a.

**Comment 4:** Figure 5: The authors claim that they randomly chose six tDMRs to subject them to independent methylation analysis using bisulphide sequencing. The chromosome 6 coordinates are given in the legend to Figure 5. This is a region contains no known gene. The authors should give the location of the analysed region within the HLA-complex and its relevance.

**Response 4:** To avoid any selection bias as part of the validation process, the six tDMRs analysed by an independent method (bisulphite DNA sequencing) were chosen randomly, irrespective of their functional relevance. However, to address this comment, we have revised Fig 5 which now shows all six tDMRs (not just one example as in the original Fig 5). By chance, the six tDMRs covered both coding and non-coding regions and their location are indicated in the revised figure. See also response 2 to reviewer 2.

**Comment 5:** Table 2 and Figure 7: The authors speak of the complement region, when they mean the C4 genes (C4A and C4B). This is confusing. The 18 tDMRs 14 to 30 and 12 map to the C4 genes. But I count 21 tDMRs (vertical black line) in Figure 7 within the complement region and also tDMRs in Table 2 map to the complement genes, which include C2. If the authors count differently, they should mark or show the vertical black lines representing the tDMR mapping to the complement region.

**Response 5:** We agree and have revised Figure 7 accordingly. Boundaries now delimit the 18 tDMRs associated with the C4A and C4B genes.

**Comment 6:** The authors concentrate in their final experiments on the complement C4 genes. This is a highly polymorphic region with regard to gene number, as two to seven C4 genes may be present in a diploid human genome. In the discussion they state the copy number variation but do not discuss whether the tissue and cell types used in this study differ with regard to the C4 gene set. The C4 gene content of DNAs analysed should be known. If the C4 gene number is the same in all DNAs, the authors should state this. Otherwise they should discuss whether different gene numbers would affect their results.

**Response 6:** This comment essentially refers to possible effects CNVs (such as C4 genes) can have with respect to sensitivity and specificity. Sensitivity is not an issue in this case because the experimental design (analysis of MeDIP-enriched versus total DNA) normalizes for the genotype of the sample DNA and is therefore independent of the genotype of the reference DNA. Therefore, the actual number of C4 genes was not determined for the sample DNAs analyzed here. Specificity is an issue but, because of the high (>99%) sequence identity between C4 A and B genes, cannot be resolved using an array of 2 Kb resolution. An explanation to this effect has been added to the discussion on page 18.

**Reviewer’s report 2**

**Comment 1:** As for any new microarray platforms, it would be nice to know the reproducibilities between technical replicates, dye swaps, and the "on-board" replicates (some r2 values should suffice).

**Response 1:** To clarify this point, we have added a new supplementary Figure 1 and the following sentence on page 13. "Control hybridizations assessing biological replicates ($R^2 > 0.97$), dye-swaps ($R^2 > 0.72$) and LM-PCR ($R^2 > 0.88$) showed that any bias introduced by these factors was within an acceptable range (supplementary Figure 1)."
Comment 2: The authors bisulfite sequenced 6 tDMR-containing tiles for validation, but this does not offer an estimate for the sensitivity. I suggest that the authors bisulfite sequence a few tiles found to be unmethylated, to see if they are indeed methylation-free.

Response 2: We are aware of the limitation that MeDIP enrichment data and bisulphite sequencing data cannot yet be directly compared to accurately determine the sensitivity the MeDIP array method. To address this issue, we have revised the corresponding part of the discussion to: “The on-going development of a novel algorithm employing a Bayesian de-convolution strategy to normalize MeDIP array data for CpG density is likely to overcome this current limitation in the near future (T. Down et al., personal communication).” In the meantime, we feel that inclusion of the generated bisulphite data are at least of confirmatory value and have revised Figure 5 to include all generated bisulphite data, including of unmethylated regions. The corresponding text on page 15 has been revised to: “For validation, we randomly selected six tDMRs (irrespective of their genomic functionality) and subjected them to independent methylation analysis using bisulphite DNA sequencing. Figure 5 shows the genomic locations of the six tDMRs (a), their methylation status based on comparison of their respective MeDIP array profiles (b) and their absolute methylation values based on bisulphite sequencing (c). In all six cases, the bisulphite sequencing results were consistent with the array data, indicating that that the array is suitable for the identification of tDMRs.” See also response 4 to reviewer 1.

Comment 3: A number of control tiles were included on the microarray but it was not clear to me how (if) they were used in this study?

Response 3: Although only the Cy3 controls were actually used for the analysis described here, we feel that it would be helpful for potential future users to include a brief description of all controls on the array. We attempted to anticipate as many controls as possible for the various studies that we envision can be carried out with the array. To address this point, we have added the following sentence on page 6: “Except for the Cy3 spots, none of other controls was required used for the analysis described here but may be useful for other types of analyses.”

Comment 4: The existence of tiles that don’t match expected genomic sequences is of concern. Although the number is small (7 out of 240; 3%), these tiles are likely amplified from elsewhere in the genome and could offer false information in downstream analysis. Would it be too expensive to sequence all the tiles, and note the “bad” ones to users?

Response 4: Like most arrays, our array is not perfect and, therefore, we attempted to sequence all tiles as suggested by the reviewer. As our tiles include repeat sequences, we further attempted to sequence each tile with both forward and reverse primers (resulting in ~3,500 sequencing reactions in total) to maximize unambiguous mapping of the tiles to the MHC. Unfortunately, the sequencing failed due to various mixups probably caused at the time by my move from Sanger to UCL where I currently do not have the resources in place to sequence on that scale. Therefore and as the number of potentially non-informative tiles can be expected to be quite low (~3% based on our original but partial QC), we hope that the following revision of the corresponding section on page 7 will be acceptable to address this point. The revised text reads: “Resequencing of 240 probes (15 % of total) identified 7 probes that failed to match to the expected reference sequences. Aliquots of all probes can be made available upon request for further QC analysis. From this partial analysis, we extrapolate that about 97% of the probes are correct and should be informative.”

Comment 5: It was not clear where the array definition files and raw data files were deposited to.

Response 5: All array data have been deposited in ArrayExpress on Thursday 15th May 2008 and the following sentence has been added on page 11: “The array data described here have been deposited in ArrayExpress under accession number ...”. We will advise of the actual accession number as soon as we receive it and will insert it into the manuscript at the proof stage at the very latest.

Comment 6: Cytosine DNA methylation is one type of DNA methylation, and it does occur in sequence contexts other than CG in some organisms (e.g., plants). The authors might want to clarify this point in the introduction.
Response 6: A brief description of non-CpG DNA methylation has been added to the introduction on page 3.

Comment 7: It was not clear which specific Dynabeads were used (e.g., protein A? G?).

Response 7: M280 Sheep anti-Mouse IgG. This information is now reported in Methods (page 8)

Comment 8: It appears that the tDMRs are enriched for potential regulatory regions, but it is difficult to judge - what is the background level (random expectations) and is the enrichment statistically significant?

Response 8: In response to this comment, we have calculated P values to identify features of statistically significant enrichment in the tDMRs. The additional analysis is described in a new section in Methods on page 11 under the header “tDMR feature analysis” and is discussed on page 16. Table 2 and its legend have been revised to indicate where enrichment is statistically significant (P < 0.05).