NIMBL
Numerical Identification of Methylation Biomarker Lists

User Guide v0.1.0

Frank Wessely[1]

[1] *School of Veterinary Medicine and Science, University of Nottingham, UK*

January 17, 2013

Contact
e-mail: svxfw@nottingham.ac.uk

Contents

1 General information

NIMBL is a software package written in Matlab to perform methylation analysis of the Infinium methylation arrays. NIMBL identifies differentially methylated sites and corresponding genes between two groups of samples. The software can compare results obtained from other methods for differential methylation analysis. NIMBL can also be used to visualise and extract methylation profiles of genes or chromosomal regions of interest within their genomic context.

Availability:
The NIMBL package, a publicly available test dataset and various annotation files can be downloaded from:
https://sites.google.com/site/emesbioinformatics/group-software

Dependencies:
NIMBL requires Matlab and the Statistics Toolbox, which is usually part of the installed Matlab version. Some parts of NIMBL (NIMBL-gene, see Section 4.3) require the Bioinformatics Toolbox. NIMBL was implemented and tested using Matlab release version R2010b (7.11) and R2011b (7.13) Linux 64-bit.

Matlab version R2012a (7.14) is not able to load larger datasets using the `importdata` function. This bug has been reported by MathWorks and is fixed in version 8.0 (R2012b). This is the description provided for bug 826575:
When reading a large text file using the Import Wizard or the importdata function, data is truncated after reading 0x3FFFFFFF (1073741823) bytes. This problem occurs only when the text file contains multi-byte encoded data.

Installation notes:
Download and unpack the NIMBL source code, the NIMBL annotation files and optionally the test dataset to any directory. Open Matlab and add this directory including all its subfolders to the Matlab search path (File, Set Path, Add with Subfolders). Alternatively, relevant directories can be added separately (File, Set Path, Add Folder). Only the latest source code (NIMBL_Scripts_v0.1.0) should be on the Matlab search path.

Running NIMBL:
NIMBL is run by calling one of its four modules (see Section 3). For example, to run the test dataset type this command in the Command Window: `nimbl('example_input');`
The file called "example_input.m" is a Matlab script specifying all the parameters that are necessary to run NIMBL analysing the test dataset. This example input file should be used as a template to set up a NIMBL run of the methylation dataset analysed. All parameters are explained in Section 3.1 and are annotated within this file, which allows a quick set-up even without prior knowledge of Matlab. All output files produced by the modules will be written

to the current working directory of Matlab. Previous files with the same file name will be overwritten with each run of NIMBL. Prefixes can be used for all file names of output files to avoid overwriting. Using Windows as the operating system might require that open PDF files need to be closed to allow overwriting.

1.1 Files within the NIMBL package

`NIMBL_Scripts_v0.1.0:`
This folder contains all Matlab scripts (.m files) of NIMBL_v0.1.0. It also contains an example input script (`example_input.m`) to analyse the test dataset and should be used as a template to run NIMBL.

`NIMBL_Annotation:`
This folder contains all annotation data necessary for NIMBL (except the genome file, see below).

`Infinium_27k_annotation.txt`
The annotation file for the Illumina HumanMethylation27 BeadChip (27k array) is based on merged information from two annotation files from Illumina:
'Illumina_HumanMethylation27-v1.2_Annotation.xls' and
'HumanMethylation27_270596_v.1.2.csv' (see Section 2.2).

`Infinium_450k_annotation.txt`
The annotation file for the Illumina HumanMethylation450 BeadChip (450k array) is based on Illumina's file:
'HumanMethylation450_15017482_v.1.1_ForExcel.csv' (see Section 2.2).

`450k_annot_design_type.txt`
File containing the 450k array IDs and a binary representation of the corresponding Infinium design type I (1) or II (0).

`450k_gene_annot.txt`
Text file containing the mapping of 450k array sites to genes and gene regions.

`hg19_chr_size.txt`
The list of the sizes of each chromosome (one row per chromosome size) corresponding to the order of the genome multi FASTA file.

`ucsc_hg19_refgene_table_raw_03112012.txt`
The gene prediction table obtained from UCSC Genome Browser (genome: human, assembly: hg19, group: Genes and Gene Prediction Tracks, track: RefSeq Genes, table:

refGene, region: genome). The rows with information about haplotype data and unplaced data (approximately 2000 rows at the end of the file) were deleted.

`ncbi_gene_info_names_03112012.txt`
: Gene information table based on 'gene_info.gz' from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/. Extracted information includes four columns: 1. 'Symbol', 2. 'Synonyms', 3. 'LocusTag' and 4. 'GeneID'.

```
awk '$1==9606{print $3"\t"$5"\t"$4"\t"$2}' gene_info \
> gene_info_names_03112012.txt
```

`refseq_removed_records_release37_to_55_human_NM_NR.txt`
: Information about removed records based on multiple archived files from: ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/. Scanned for human entries and NM and NR accessions. Extracted information includes three columns: 1. accession (version removed), 2. refseq status and 3. removed status.

```
awk -F $'\t' '$1==9606 && $3~/N[M|R]/ { split($3, s, "."); $3 = s[1]; \
print $3"\t"$6"\t"$8 }' removed_records_release37_to_55.txt \
> refseq_removed_records_release37_to_55_human_NM_NR.txt
```

`ucsc_hg19_refgene_table.txt`
: The processed UCSC gene table with an updated column of gene symbols (original column 'name2'). This file is the output file of the additional pre-processing script `preprocess_gene_info` (see Section 4.3) This file is used within NIMBL-gene to extract location information of user-specified genes.

`preprocess_gene_info_report.txt`
: The report of the preprocessing script obtained via `preprocess_gene_info('example_input');`

Additionally, on our website a human genome file obtained from UCSC Genome Browser can be downloaded, which is necessary for using the module NIMBL-gene (currently works only with 450k data). The unpacked FASTA file could be placed after download within the annotation folder of NIMBL. Otherwise ensure that the file is on the search path of Matlab.

`hg19_genome_repeat_lower.fa`
: A multi FASTA file of the human genome. The 24 single FASTA files obtained from the UCSC Browser of each chromosome sequence, where repeats are indicated by lower case letters, are concatenated in the order 1,2,3,...,9,10,11,12,...20,21,22,X,Y.

NIMBL_Example:

`GSE29290_BC_16_samples.txt`

The test dataset that is available from Gene Expression Omnibus (GEO) database (GEO Accession GSE29290). This dataset consists of eight breast tumour samples and eight normal breast tissue samples (Dedeurwaerder et al., 2011), for which methylation levels were measured with the Illumina HumanMethylation450 BeadChip. The file includes the methylation levels as beta values and corresponding detection P-values to estimate the quality of the measurement.

`GSE29290_BC_16_samples_nonorm_nopeak_480917.txt`

The methylation dataset used for differential methylation analysis by NIMBL and IMA (Illumina Methylation Analyzer, (Wang et al., 2012)). Array sites with one or more of the 16 samples having a detection p-value of less than 0.05 and sites with beta values were excluded. This dataset contains a total of 480917 measurements.

`nimbl_example_BC_d0.1_m_0_8_diff_methylation_IDs.txt`

The list of 1347 differentially methylated sites identified by NIMBL for the test dataset (minimal intergroup beta value distance 0.1, no samples masked).

`nimbl_example_BC_d0.1_m_2_8_diff_methylation_IDs.txt`

The list of 15047 differentially methylated sites identified by NIMBL for the test dataset (minimal intergroup beta value distance 0.1, up to two samples masked).

`IMA_example_BC_wilcoxon_BH_P_0.05_IDs.txt`

The list of 32184 differentially methylated sites identified by IMA using the wilcoxon option. The Benjamini Hochberg (BH) procedure was used for multiple testing correction and the cut-off for adjusted p-values was set to 0.05.

`IMA_example_BC_limma_BH_P_0.05_IDs.txt`

The list of 35629 differentially methylated sites identified by IMA using the limma option. The Benjamini Hochberg (BH) procedure was used for multiple testing correction and the cut-off for adjusted p-values was set to 0.05.

2 Input data files

2.1 Methylation data

The input methylation profiles of all samples must be provided as a text file. Columns in the text file are TAB-delimited or the user specifies another delimiter. The input file must contain for each sample:

- beta values or signal intensities to calculate the beta values
- detection p-values

The column order of the samples specifying the beta values must correspond to the column order of the samples specifying the p-values. For example, the first column (from left) of beta values and the first column of detection p-values must be from the same sample. The beta value columns must contain the keyword `'beta'` within the column name, whereas the p-value columns must contain the keyword `'pval', 'p-val' or 'p val'`. Missing values must not be indicated by the keyword 'NULL'; 'NaN' or empty values should be used at these positions. Sample names are extracted from the first row of the input file, where they are separated from the keywords by a certain delimiter (see Section 3.1.1). One column must contain the unique probe identifier of each measurement. This column and any additional annotation data within the file must be on the left-hand side of the numerical values. Additional annotation columns of array sites are not further used within NIMBL. Including only the necessary information (unique identifiers, beta values and p-values) guarantees the fastest reading performance. The example dataset (`GSE29290_BC_16_samples.txt`) is an example of an input file with minimum input data. The methylation input file can contain any number of measurements (i.e. number of rows).

2.2 Annotation data

Array sites are annotated by a text file provided by the user or by the annotation files provided within the `NIMBL_Annotation` folder (see Section 1.1). Both annotation files of the two array types contain, besides the information provided by Illumina, an additional column called 'CPG Region UCSC'. This column specifies the location of each CpG site compared to its nearest CpG island (CGI) by providing a number for each CpG site:

1: within CGI

2: N-shore (distance <= 2kb upstream of CGI start)

3: S-Shore (distance <= 2kb downstream of CGI end)

4: N-Shelf (2kb < distance <= 4kb upstream of CGI start)

5: S-Shelf (2kb $<$ distance $<=$ 4kb downstream of CGI end)

6: 'deep sea' ($>$ 4kb away from any CGI)

For the 450k array Illumina's column called 'Relation_to_UCSC_CpG_Island' was translated into the corresponding numerical values (1...6). This information in turn is based on Illumina's reported CpG islands ('UCSC_CpG_Islands_Name'). For the 27k array CGIs were recalculated by using the method used by the UCSC genome browser and are reported in the column 'CPG_ISLAND_LOCATIONS hg18 UCSC algo'. Calculations were based on human genome version hg18. Settings were: length $>=$ 200, GC content $>=$ 50%, O/E ratio $>=$ 0.6, cpgscore $=$ 17 (used to increment the running sum). Illumina's original CGI annotation of the 27k array is given in the column 'CPG_ISLAND_LOCATIONS'.

The user can provide any other annotation information to select a subset of array sites. For example, the default annotation files contain an additional column called 'user annotation' that can be used in a similar way as the CpG region column. Incorporating additional selection criteria might require a modification of the script `select_probes.m`.

2.3 Compatibility of input data files

For each array site given in the methylation file there must be a corresponding row within the annotation file. If the methylation file does not contain the full set of array sites for the chosen array type (e.g. due to any user-defined pre-selection of array sites independent of NIMBL), annotation is only retained for the sites provided within the methylation input file. The NIMBL-gene module (see Section 4.3) only considers sites that are given in the methylation input file. This means that the total number of array sites corresponding to a gene or chromosomal region might be higher. However, the gene output tables, e.g. `nimbl_gene_table.txt` (see Section 4.2.1), still report the total number of array sites corresponding to a gene (sites not given in the input file are treated as excluded sites).

3 Running NIMBL

NIMBL consists of four main modules:

- NIMBL-qc (`nimbl_qc('example_input');`)
- NIMBL (`nimbl('example_input');`)
- NIMBL-gene (`nimbl_gene('example_input');`)
- NIMBL-compare (`nimbl_compare('example_input');`)

NIMBL is used by calling one of these modules with the name of the input script as the only parameter. To do so type the name of the module with the relevant input script name in the Command Window (as exemplified in the brackets above). Each output file generated contains the module name as a prefix (or any prefix specified by the user, see Section 3.1.7). Each module produces a report (TAB-delimited text file) to give an overview of main settings and obtained results. All output text files are TAB-delimited and all output plots are PDF files. The parameters of the input script are explained in Section 3.1, whereas the individual modules and their output files are explained in Section 4.

3.1 Input script

The example input script (`example_input.m`) should be used as a template. The names of the parameters must not be changed.

3.1.1 Parameters for input files

1.1 `input_file`

File name for methylation input data. If the file is not in the current working directory of Matlab, the search path must be changed to include the directory of the input file (File, Set Path, Add Folder). Please see Section 2.1 for the required file format. If `input_file` is not changed during successive runs of NIMBL, this file is not imported again. That means if the file is modified but the file name is not changed during the current Matlab session, type `clear read_infinium` in the Command Window to ensure that the updated file is imported.

1.2 `annot_file`

Optional annotation file name. To use the default files set `annot_file = '';`. If `annot_file` is not changed during successive runs of NIMBL, this file is not imported again. That means if the file is modified but the file name is not changed during the current Matlab session, type `clear read_annotation` in the Command Window to ensure that the updated file is imported.

1.3 `input_del`

The column delimiter used in the input files, default: TAB (`'\t'`).

1.4 `platform`

The Infinium platform: 27k array (1) or 450k array (2).

1.5 `sample_del`

The delimiter for sample names within columns of header of input file (e.g. `'_'` or `'.'`).

1.6 `sample_pos`

The position of sample name in column headers: before sample delimiter (1) or after (2). Examples:

- 'PRL_1 (M).pval': 'PRL_1 (M)' is the sample name (set `sample_pos = 1;` and `sample_del = '.';`)
- 'pval_mcf7.untr': 'mcf7.untr' is the sample name (set `sample_pos = 2;` and `sample_del = '_';`)

3.1.2 Parameters for quality control of samples

2.1 `pval_cut_s`

The threshold for detection p-value, values above threshold indicate low-quality measurement (e.g. 0.05).

2.2 `sample_ex`

The indices of samples to exclude from methylation analysis. The first sample of the input file has the index 1. Examples:

- Do not exclude any sample: set `sample_ex = [];`
- Exclude samples 1 and 12: set `sample_ex = [1,12];`

2.3 `qc_ex`

Choose whether samples marked for exclusion should also be excluded from the QC procedure (no: 0, yes: 1).

2.4 `kde_test`

Generate (1) or skip (0) plots of beta value distribution within NIMBL-qc. Plotting the beta value distribution of each sample might take some time. The default settings can be changed within the script `'get_beta_kde.m'`:
`kernel = 'normal'`, `width = 0.05`, `npoints = 200`

2.5 `peak_correct`

Choose whether beta values of Infinium design type II probes should be corrected. The procedue is based on the peak-based correction described in (Dedeurwaerder et al., 2011). All samples are either identically or individually corrected. For an identical correction of all samples, set `peak_correct` to:

`[0,0]` no correction of any sample (0)

`[1,1]` correct both peaks of every sample (1)

`[1,0]` correct only unmethylated peak of every sample (2)

`[0,1]` correct ony methylated of every sample (3)

To correct each sample individually specify one integer (0,1,2 or 3) for every sample. Examples:

- Do not correct any sample: set `peak_correct = [0,0];`
- Correct both peaks of samples 1...6, but correct only unmethylated peaks of samples 7...12: set `peak_correct = [1,1,1,1,1,1,2,2,2,2,2,2];`

2.6 `peak_correct_detail`

If any peaks are corrected, choose the plots of Infinium I and II distribution during quality control of NIMBL-qc. Example:

- Merge all samples in one plot: set `peak_correct_detail = 0;`
- Individual plot of each sample: set `peak_correct_detail = 1;`

2.7 `peak_correct_write`

Write the peak-based corrected beta values to a text file. The file name is based on `input_file` with the prefix 'NIMBL_peak_corrected'. If samples are excluded from the quality control procedure (`qc_ex = 1`), these are also not exported here. Choose between three options:

0 no output

1 print corrected beta values

2 print corrected beta values and p-values (generate a valid NIMBL input file)

3.1.3 Parameters for quality control and selection of array sites

3.1 `pval_cut_p`

Threshold for detection pvalue, values above threshold indicate low-quality measurement. This parameter is also used to highlight low quality measurements within the output plots of the NIMBL-gene module. Measurements above the threshold are marked with an 'x' within the circles. This NIMBL-gene option can be switched off by setting the threshold to 1.

3.2 `pval_qc_p`

Threshold in percent of accepted number of low quality measurements per array site, exclude sites above threshold (set value to 0 for most stringent QC)

3.3 `miss_beta_qc`

Fraction of maximal number of samples with missing beta value (NaN) in each group [group1, group2], values < 1. Examples:

- Exclude sites with missing beta values in any group:
  set `miss_beta_qc = [0, 0];`
- Allow 25% of samples in group1 to have a missing beta value, none such in group2:
  set `miss_beta_qc = [1/4, 0];`

3.4 `chr_sel`

Selection of chromosomes, binary vector of four chromosome groups: (1) autosomes, (2) X, (3) Y, (4) unspecified. Examples:

- Select all chromsomes and unspecified: set `chr_sel = [1,1,1,1];`
- Exclude X and Y chromosome: set `chr_sel = [1,0,0,1];`

3.5 `cpg_reg`

Select CpG region (see Section 2.2), several regions can be selected:

1 CpG island (CGI)

2 N-Shore (5' shore)

3 S-Shore (3' shore)

4 N-Shelf

5 S-Shelf

6 not related to CGI

Examples:

- Do not select for CpG region: set `cpg_reg = 0;`
- Select sites within CGIs: set `cpg_reg = 1;`
- Select sites in CGIs and shores: set `cpg_reg = [1,2,3];`

3.6 `cpg_user`

Select according to user annotation column, specify any numerical value within column, same usage as `cpg_reg`. To ignore this option set `cpg_user = 0;`

3.7 `selected_write`

Write selected array sites to a TAB-delimited text file. These are the sites for which differential methylation analysis is performed by the NIMBL core module. The name of the methylation input file (`input_file`) with the suffix 'NIMBL_selected' is used as the file name. The column header includes the keywords 'Beta' and 'Pval' as a suffix separated by the delimiter specified by `sample_del` from the original sample names. If samples are excluded from the analysis (by setting `sample_ex`), these will not be printed. If peak correction is performed within NIMBL, peak-based corrected beta values are exported, the suffix of the output file is 'NIMBL_peak_corrected_selected' and the suffix of the sample names in the column header is 'BetaPeakCorrected'. Five options are available:

0 no output

1 export cgIDs (array identifiers)

2 export cgIDs and beta values

3 export cgIDs and beta values and p-values

4 export cgIDs and beta values and p-values and array annotation from input file

3.8 `excluded_write`

Write excluded array sites to a TAB-delimited text file. These are the sites which are not considered for differential methylation analysis. The same syntax with five options as for parameter `selected_write` applies.

3.1.4 Parameters for sample groups and differential methylation

4.1 `g1` **and** `g2`

Choose two groups of samples, use sample index according to occurrence of samples in input file. The NIMBL-gene module allows to specify additional groups (`g3`, `g4`, etc; see 4.3). Examples:

- 12 samples, 4 in group1 and 4 in group2: set `g1 = [1,2,5,6];` and `g2 = [3,4,11,12];`
- 24 samples, 12 in group1 and 12 in group2: set `g1 = [1:6,13:18];` (equivalent to `g1 = [1,2,3,4,5,6,13,14,15,16,17,18];`) and `g2 = [7:12,19:24];` (equivalent to `g2 = [7,8,9,10,11,12,19,20,21,22,23,24];`)

4.2 `group_names`

Names of the two groups. The NIMBL-gene module allows to specify additional groups and names (see 4.3). Example: name group1 is 'Treatment', name group2 is 'Control', set `group_names = {'Treatment', 'Control'}`.

4.3 `beta_dist`

Minimal beta value distance between non-overlapping groups.

4.4 `mask_frac`

Fraction of maximum number of masked samples in each group [group1, group2], values < 1. Examples:

- Do not mask any sample: set `mask_frac = [0,0];`
- Mask up to 25% of samples in group1, no masking in group2: set `mask_frac = [1/4,0];`
- Mask up to 3 out of 12 samples in group1 and group2: set `mask_frac = [3/12,3/12];`

4.5 `limit1` and `limit2`

Optional lower and upper limits on median beta values of each group [lower, upper]. Examples:

- No constraints for any group: set `limit1 = [0,1];` and `limit2 = [0,1];`
- Hypomethylation group1 and hypermethylation group2: set `limit1 = [0,0.3]; limit2 = [0.7,1];`

4.6 `hits`

Choose a subset of identified differentially methylated sites according to their rank for plotting methylation levels of group1 and group2. This subset can be individual sites ([rank1; rank2; rank3]) or a range of of sites [rank start, rank end]. Examples:

- Plot top 50 sites: set `hits = [1,50];`
- Plot 5 individual sites: set `hits = [10;11;12;80;81];`

4.7 `flag_mask`

Flag beta values of masked samples with NaN in output text file of differentially methylated sites identified (set `flag_mask = 1;`). If beta values of masked sample are not flagged (set `flag_mask = 0;`), NaN fields in the output file correspond to missing beta values.

4.8 `boxplot_write`

Print boxplots of overall methylation distribution of each sample within group1 and group2. Three options are available:

0 no boxplot

1 boxplot with samples in original order

2 boxplot with group1 samples left and group2 samples right

3.1.5 Parameters for gene or region-specific analysis (NIMBL-gene)

5.1 `nimbl_gene_input`

List of genes or chromosomal regions of interest, or file name of file with listed genes or regions (one row per input value). Genes can be specified by general gene names or RefSeq accession IDs (NM and NR accessions). Specific accessions can be chosen by specifying the transcription start site after the ID with a ';' as delimiter. Chromosomal regions are specified by four or five values separated by ';' in this order: 0. optional name for this region; 1. chromosome; 2. strand ('+' or '-'); 3. start coordinate; 4. end coordinate. Start and end coordinates are 1-based. Examples:

- Read input from file 'nimbl_detail_list.txt':
  set `nimbl_gene_input = 'nimbl_detail_list.txt';`
- Analyse one gene:
  set `nimbl_gene_input = {'CHAD'};`
- Analyse two genes:
  set `nimbl_gene_input = {'SLC38A2', 'NM_032785'};`
- Analyse one specific gene transcript:
  set `nimbl_gene_input = {'NM_032785;1234567'};`
- Analyse one region:
  set `nimbl_gene_input = {'14;+;78227173;78236085'};`
- Analyse one region called 'my Region':
  set `nimbl_gene_input = {'my Region;X;-;123345678;12346678'};`

5.2 `extra_upstream`

Include that many extra bases upstream of transcription start site (TSS, txS) of the genes specified. For example, to include 2,000 extra bases set `extra_upstream = 2000;`

5.3 `extra_downstream`

Include that many extra bases downstream of transcription end site (txE) of the genes specified. For example, to include 200 extra bases set `extra_downstream = 200;`

5.4 `align_probes`

Switch to perform alignment of probes and generating multi FASTA files of aligned probes to specified genes or chromosomal regions. Alignments are performed with `align_probes = 1;` and skipped with `align_probes = 0;`.

5.5 `zoom_in`

Create additional methylation overview plots, profile plots and text files with methylation levels by selecting a certain plot region [start position, end position]. This might be useful if the genomic sequence is very long. If genes are specified, the start of the plot (position 1) is defined by number of extra bases upstream of the transcriptional start site. The suffix '_zoom' within the output file names indicates the usage of this option. Examples:

- Disable zoom plot: set `zoom_in = 0;`
- Show first 3,000 bases: set `zoom_in = 3000;`
- Display region from position 2,000 to 8,000: set `zoom_in = [2000, 8000];`

5.6 `label_sites`

Switch to show labels of identified array sites on the x-axis of the methylation overview plot. The labels represent the positions of the array sites within the specified genomic sequence (5' to 3'). These labels correspond to the order of array sites within the alignment file. Show labels with `label_sites = 1;` and do not show them with `label_sites = 0;`.

5.7 `methyl_profile_write`

Control the text file output of methylation levels of array sites, which were found within the specified genomic region. Three options are available:

0 no output

1 array site ID, beta values and p-values

2 all information from probe annotation file, beta values and p-values

5.8 `sites_extra`

File name of specific array sites (specified by array IDs, i.e. mainly cgIDs) given in one or more columns/groups (TAB-delimited). Each column of array IDs must have a header within the text file. Specified array sites are highlighted by colored stems within the two output plots of NIMBL-gene. If only one column is provided the stems are plotted in green, otherwise colors are obtained according to the color scheme of the script `get_color_code`. The legend of the plots contains the header of the columns and the number of sites found within the specified column (only columns with at least one member are shown within the legend). If sites belong to more than one column, only the last membership (i.e. the most right column) is plotted and counted. It can be that Matlab clips the names within the legend box or the layout within the box is not proper. It seems that using capital letters for names within the legend causes such layout problems. If corresponding methylation levels are written to a text file (chosen by `methyl_profile_write`), a binary matrix on the right-hand side indicates which sites are found within each group. For example, sites can be highlighted that were found as differentially methylated by NIMBL or any other method, or sites that were excluded from differential methylation analysis.

5.9 `refgene_ucsc_table`

File name of raw UCSC refGene table (see Section 4.3).

5.10 `gene_info_ncbi_table`

File name of extracted information from NCBI 'gene_info' file, only human entries. Extracted information includes four columns: 1. 'Symbol', 2. 'Synonyms', 3. 'LocusTag' and 4. 'GeneID' (see Section 4.3).

5.11 `refseq_removed_table`

File name of removed RefSeq records from NCBI, only human entries. Extracted information includes three columns: 1. accession (version removed), 2. refseq status and 3. removed status.

3.1.6 Parameters for comparison of sites

6.1 `n`

Number of lists to compare against each other (2 or 3).

6.2 `list1`, `list2` (, `list3`)

File names of lists containing the array sites. If no full path is specified, the files need to be within the Matlab search path.

6.3 `m1`, `m2` (, `m3`)

Names of the two or three input lists. These names are used for figure legends and also the file names of output files, so that certain characters cannot be used within these names.

6.4 `beta_file`

File name of table with identifiers of array sites in the first column and beta values in remaining columns. Columns must be TAB-delimited and one row as header (column names) is required, which is not further used. The order of the samples must correspond to the groups chosen by `g1` and `g2`. This table represents the list of all selected array sites that were used to identify the sites given in the input lists. If no full path is specified, the file needs to be within the Matlab search path. If `beta_file` is not changed during successive runs of NIMBL, this file is not imported again. That means if the file is modified but the file name is not changed during the current Matlab session, type `clear nimbl_compare` in the Command Window to ensure that the file is read again.

6.5 `f_on`

Optional post-filter lists based on mean or median beta value difference between group1 and group2. Binary vector of length `n`. Examples:

- Post-filter all 3 input lists: set `f_on = [1,1,1];`

- Post-filter only the first list of 3 input lists. set `f_on = [1,0,0];`
- No post-filter of any list: set `f_on = [0, 0];` (2 input lists) or`f_on = [0, 0, 0];` (3 input lists)

6.6 `filter_method`
Post-filter based on mean (1) or median (2) beta value difference between groups.

6.7 `f`
The minimum mean or median beta value difference between groups used to post-filter array sites.

6.8 `m`
Plot mean (2) or median (2) beta values of groups against each other in the overview scatter plot.

6.9 `gene_table_sep`
Optional output of the gene information table of every gene set that is unique to one of the input lists or any combination. Set `gene_table_sep = 0;` to disable it. Enabling this option (`gene_table_sep = 0;`) generates up to 3 (`n=2`) or 7 (`n=7`) additional text files.

3.1.7 Parameters for output file names

7.1 `nimbl_qc_prefix`
Prefix for files produced by NIMBL-qc module (default: nimbl_qc).

7.2 `nimbl_prefix`
Prefix for files produced by NIMBL module (default: nimbl).

7.3 `nimbl_gene_prefix`
Prefix for files produced by NIMBL-gene module (default: nimbl_gene).

7.4 `nimbl_comp_prefix`
Prefix for files produced by NIMBL-compare module (default: nimbl_compare).

4 Modules

4.1 NIMBL-qc

This module allows for a basic quality assessment of samples. Several plots are generated to visualize the sample quality. Deviation from the expected distribution is largely related to the detection p-values, where an increase in number of measurements with low confidence of methylation accuracy (e.g. detection p-value > 0.05) is reflected by a significant deviation of beta value distribution. A Kolmogorov-Smirnov test (see Section 4.1.1) is performed to assist the identification of low quality samples which may influence downstream analysis and these can be then excluded from further analysis.

4.1.1 Kolmogorov-Smirnov test

The module implements the two-sample Kolmogorov-Smirnov (KS) test (as provided by Matlab) to find any aberrant beta value distributions. This non-parametric test tests whether two samples are from the same distribution (null hypothesis) by comparing their cumulative frequencies. As a reference sample, we chose the median value of all samples from the input beta values. Each sample is compared to this reference median sample, which means that the number of tests corresponds to the number of samples. The observed KS test statistic represents the maximum difference between the cumulative distribution functions of the sample and the reference. If the calculated (asymptotic) P-value is smaller than a chosen significance level (Bonferroni-corrected by the number of test performed), the null hypothesis is rejected. This indicates a different distribution from the median reference.

However, applying the KS test to the complete Infinium data with large sample sizes (approximately 30000 or 480000 for the 27k and 450k arrays respectively) increases the power of the test, which leads in almost all cases to the rejection of the null hypothesis. This means that even small differences between the samples and the median reference become statistically significant. Therefore, we implemented a heuristic that samples a fraction from each original sample and the median reference. A histogram of each sample (and the reference) is computed with 100 bins and a fraction of associated measurements is chosen (default is 1/100 and 1/1000 for the 27k and 450k arrays respectively). The KS test results are given within the report of the module. If the null hypothesis is rejected for any sample, beta value distribution plots highlight these samples.

Please note that this heuristic does not answer which samples need to be excluded from further analysis. The aim of the heuristic is to help the user to spot samples showing a distinct beta value distribution. The aberrant distribution of samples might be explained by the underlying biological experiment itself. However, our experience with the heuristic is that even if very different samples (e.g. normal vs cancer) are included in a dataset, only strong outliers are detected. For finding truly problematic samples, we highly recommend to incorporate the other quality control plots of the module, especially the plot showing the number of low quality measurements of each sample, which is measured by detection p-values above a certain threshold (e.g. 0.05). The NIMBL-qc module offers only a basic global quality control. Other tools, for example, the R packages lumi (Du et al., 2008), methylumi or HumanMeth27QCReport (Mancuso et al., 2011), offer comprehensive quality control procedures and pre-processing steps. Using these tools is recommended, especially if some samples show a different behaviour based on the global tests performed by the NIMBL-qc module.

4.1.2 Output files

`nimbl_qc_report.txt`

Summary information of module results including information about samples obtained from input file, descriptive statistics about missing beta values and high detection p-values and Kolmogorov-Smirnov test results.

`nimbl_qc_beta_distribution.pdf`

Beta value distribution (kernel smoothing density estimate) of all samples.

`nimbl_qc_pval_total.pdf`

Bar plot of the number of low quality measurements of each sample.

`nimbl_qc_pval_percent.pdf`

Bar plot of the percent values of low quality measurements of each sample.

`nimbl_qc_miss_beta.pdf`

Bar plot of the number of missing beta values of each sample.

`nimbl_qc_beta_CDF.pdf`

Empirical cumulative distribution function plot of each sample.

`nimbl_qc_beta_CDF_subset_KS.pdf`

Empirical cumulative distribution function plot of each reduced sample (see Section 4.1.1).

`nimbl_qc_beta_CDF_KS.pdf`

If the KS-test detects any aberrant sample, the empirical cumulative distribution function plot highlighting detected samples.

`nimbl_qc_beta_distribution_KS.pdf`

If the KS-test detects any aberrant samples, the beta value distribution (kernel smoothing density estimate) of all samples highlighting detected samples.

`nimbl_qc_design_type_beta_dist.pdf`

If peak-based correction of beta values of Infinium design type II probes is performed, the distribution plot of beta values of all samples separated by Infinium design type I and II before correction.

`nimbl_qc_design_type_beta_dist_corrected.pdf`

If peak-based correction is performed the distribution plot of beta values of all samples separated by Infinium design type I and II after correction.

4.2 NIMBL

This module identifies differentially methylated sites and corresponding genes. Detection of differential methylation requires the selection of any two groups of samples (see Section 3.1.4). Array sites with measurements of low confidence in a number of samples, can be excluded from differential methylation analysis. Differentially methylated sites are identified as sites with the largest difference in methylation levels between the two groups. The user can control this by specifying a minimum beta value distance between non-overlapping groups (`beta_dist`). There is the option to mask a proportion of samples which are highly heterogeneous (`mask_frac`). Samples which are most distant from the median beta value of each group are successively removed until the minimum discrimination of methylation levels is fulfilled or the maximum number of masked samples is reached. The list of potential biomarkers is ranked by calculating a score based on the inter-group and intra-group variability:

$$score = beta\_val_{dist} - (median_{diff} - beta\_val_{dist}) \tag{1}$$

where $beta\_val_{dist}$ is the distance in beta values between non-overlapping groups and $median_{diff}$ is the absolute difference of the medians of each group. Higher discrimination between groups and lower variability within groups yields a higher score. The ranked list of potential biomarkers is visualised in a summary plot and reported along with annotation information in a text file. Further constraints can be imposed by imposing limits on the median beta values of each group (`limit1` and `limit2`). The module generates a comprehensive table of genes associated with the sites identified. This gene-centric analysis allows the rapid investigation of enrichment of differentially methylated sites within the genes and their corresponding regions.

4.2.1 Output files

`nimbl_report.txt`

Summary information of module results including information about array sites selected for differential methylation analysis (see Section 3.1.3), groups compared, differential methylation settings and results, information about masking of samples and tables showing the variation of beta values of array sites of each group.

`nimbl_beta_distribution_selected_sites.pdf`
: Beta value distribution (kernel smoothing density estimate) of all samples based on array sites used for differential methylation analysis.

`nimbl_groups_boxplots.pdf`
: Boxplots of each sample based on the array sites selected. The two sample groups are highlighted by two colors (red and blue). Samples can be plotted in the original sample order or group1 samples are plotted on the left-hand side and group2 samples on the right-hand side.

`nimbl_diff_methylation.txt`
: Ranked list of differentially methylated sites with complete annotation and information about the number of masked samples.

`nimbl_diff_methylation.pdf`
: Methylation levels of a subset (specified by `hits`) of differentially methylated sites highlighting methylation levels of group1 versus group2.

`nimbl_diff_methylation_IDs.txt`
: List of unique identifiers of differentially methylated sites.

`nimbl_gene_table.txt`
: Text file of differentially methylated genes and their enrichment of differentially methylated sites in their gene regions. This analysis is currently only available for 450k array data. The gene regions obtained from the 450k annotation file are: TSS1500 (region up to 1500bp upstream of transcription start site - TSS), TSS200, 5'UTR, first exon, gene body, 3'UTR and promoter (summary of TSS1500 + TSS200 + 5'UTR). Note that array sites can have multiple annotations for gene regions. In these cases sites are counted for every region once. The table is sorted according to the number of sites found as differentially methylated and the fraction of these sites within all sites used for differential methylation analysis.

4.3 NIMBL-gene

This module allows the detailed examination of user defined genes or chromosomal regions of interest. It is the main module to visualise methylation levels of any genomic region. Input is single or multiple gene symbols, RefSeq accession IDs or chromosomal regions, which are specified in the input script or listed in separate text file, see Section 3.1.5 for an explanation of correct syntax to specify input values. Genomic location information of genes (chromosome, transcription start site, etc.) is extracted from a pre-processed version of the UCSC hg19 refgene table ('ucsc_hg19_refgene_table.txt'). The pre-processing compares and expands the UCSC gene symbols provided within the refgene table (column 'name2') with NCBI gene symbols and synonyms extracted from 'gene_info.gz'. Updated versions of these files can be processed with the script `preprocess_gene_info`, which is within the NIMBL scripts folder. This script also provides a report based on the comparison of gene annotation from the array annotation file and UCSC/NCBI annotation. The NIMBL annotation folder contains the report file based on the UCSC and NCBI annotation files specified within the example input script.

The processed UCSC refgene table ('ucsc_hg19_refgene_table.txt') provides within the original column 'name2' all identified and unique gene symbols and synonyms separated by '|'. This increases the probability of finding genes specified as input values to NIMBL-gene. User-specified gene symbols are found within this column by exact matches. If several RefSeq accessions are found for one gene symbol, the longest transcript is used for further analysis. The NIMBL-gene report file provides an overview table of all matches of gene symbols.

The search of corresponding array sites is only based on genomic location of measurements (i.e. 'mapinfo' column within the annotation file), hence array sites are found independently of the (potentially outdated) gene annotation of the array probes. Genomic information is used to create methylation overview plots of each gene or chromosomal region, where methylation measurements are plotted according to their genomic location. Additional groups can be specified to highlight specific samples within methylation overview and profile plots. To do so specify additional group parameters `g3`, `g4`, etc. The number and the index of these groups must be monotonically increasing. For example, to consider a total of four groups of samples specify `g1` to `g4` (see Section 3.1.4). To keep the color scheme of groups consistent (group1 samples always in red, group2 sample always in blue, etc.), empty groups can be specified (e.g. `g3=[];` while g4 can still be used e.g. `g4=[9,13:16];`). The color scheme is fixed for up to seven groups and can be changed within the script `get_color_code`.

Optional alignment files in multi FASTA format, which contain the DNA sequence of the specified gene or chromosomal region and the aligned probe sequences, provide information of the exact sequence context of array measurements. To ensure that the sequence of array probes at the edge of genomic regions can be completely aligned to the genomic sequence, 50 extra bases are added to the specified genomic region.

4.3.1 Output files

`nimbl_gene_report.txt`

Summary information of module results including information about the genes found in the RefSeq table from UCSC Genome Browser and their corresponding genomic information.

`nimbl_gene_methylation_overview_NAME.pdf`
: Overview of all and measured CpG sites of the gene or chromosomal region 'NAME' according to position in genomic sequence (5' to 3'). All CpG sites within the sequence are displayed as small stems below the x-axis. Low-quality measurements (based on the threshold specified by `pval_cut_p`) are indicated with 'x'. Specific array sites can be color-highlighted by specifying an input file of array IDs via `sites_extra`.

`nimbl_gene_methylation_overview_NAME_zoom.pdf`
: Zoomed-in version of the methylation overview plot.

`nimbl_gene_methylation_profile_NAME.pdf`
: Methylation levels of the array sites within genomic region in genomic order (5' - 3') but not regarding actual genomic distance as in overview plots. Low-quality measurements (based on the threshold specified by `pval_cut_p`) are indicated with 'x'. Specific array sites can be color-highlighted by specifying an input file of array IDs via `sites_extra`.

`nimbl_gene_methylation_profile_NAME_zoom.pdf`
: Zoomed-in version of the methylation profile plot.

`nimbl_gene_methylation_profile_NAME.txt`
: Text file with corresponding array IDs, beta values and detection P-values and optionally additional probe annotation. If extra array sites are specified by `sites_extra` to be highlighted, a binary matrix is appended to the right-hand side of the output table. It reports which array site is found in which column/group of the input file.

`nimbl_gene_methylation_profile_NAME_zoom.txt`
: Zoomed-in version of the output text file.

`nimbl_gene_probe_alignment_NAME.fa`
: Multi FASTA sequence file with the DNA sequence (5'-3'), aligned array probes, CpG sites in total and CpG sites measured on the array. Please use a sequence viewer to explore this file, for example, seaview (Gouy et al., 2010).

4.4 NIMBL-compare

This module can be used to compare two or three lists of sites identified by any method or by different settings of one method (see Section 3.1.6). The comparison is performed both on a site and gene level. The output gene information tables can be used to identify candidate genes that are common or unique between the input lists or their combinations. It is also possible to use this module to generate the gene information table based on only one input list. Simply compare the sites of interest to an empty text file as the second list.

4.4.1 Output files

`nimbl_compare_report.txt`
: Summary information of module results including information about mean and median group differences, post-filter results, overlap of input lists on site and gene level.

`nimbl_compare_scatter_plot_(mean|median)_L1_L2(_L3).pdf`
: Overview scatter plot of the overlap of all or post-filtered sites of the input lists L1, L2 (, L3).

`nimbl_compare_IDs.txt`
: Lists of array sites (identifiers) unique to one of the input lists or any combination (3 or 7 combinations in total for 2 or 3 input lists respectively). Note that these lists of unique sites do not directly translate into lists of unique, since genes can be detected by multiple sites.

`nimbl_compare_genes_total.txt`
: Complete lists of genes corresponding to the each input list separately (2 or 3 columns). That means genes can be listed only once or up to 2 or 3 times.

`nimbl_compare_genes.txt`
: Lists of genes which are unique to each input list or any combination (3 or 7 columns).

`nimbl_compare_gene_table.txt`
: Comprehensive gene information table with all genes related to all sites of the input lists, reporting the number of sites provided by each input list individually or any combination. It also includes the enrichment of sites, which are reported in any input list, within the following gene regions: TSS 1500 (region up to 1500bp upstream of transcription start site - TSS), TSS200, 5'UTR, first exon, gene body, 3'UTR and promoter (TSS 1500 + TSS200 + 5'UTR). Note that array sites can have multiple annotations for gene regions. In these cases sites are counted for every region once. The table is sorted according to the number of total sites found as differentially methylated (i.e. a site given in one or more input lists) and the fraction of these sites within all sites used for differential methylation analysis. The module can also provide a similar gene table (text file) for any list of genes, which are unique to one input list or any combination of them (specified by `gene_table_sep`, see Section 3.1.6). The corresponding names of the lists or their combination (indicated by a '+') are used within the file names.

References

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**: 771–784

Du P, Kibbe WA, Lin SM (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**: 1547–1548

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**: 221–224

Mancuso FM, Montfort M, Carreras A, Alibés A, Roma G (2011) HumMeth27QCReport: an R package for quality control and primary analysis of Illumina Infinium methylation data. *BMC Res Notes* **4**: 546

Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28**: 729–730