# Supporting diagnostics and therapy BatchI: Batch effect Identification in high-throughput screening data using a dynamic programming algorithm

**Anna Papiez, Michal Marczyk, Joanna Polanska, Andrzej Polanski**

**Bioinformatics 35 (2019) 1885-1892**

**Abstract**

**Motivation:** In contemporary biological experiments, bias, which interferes with the measurements, requires attentive processing. Important sources of bias in high-throughput biological experiments are batch effects and diverse methods towards removal of batch effects have been established. These include various normalization techniques, yet many require knowledge on the number of batches and assignment of samples to batches. Only few can deal with the problem of identification of batch effect of unknown structure. For this reason, an original batch identification algorithm through dynamical programming is introduced for omics data that may be sorted on a timescale.

**Results:** BatchI algorithm is based on partitioning a series of high-throughput experiment samples into sub-series corresponding to estimated batches. The dynamic programming method is used for splitting data with maximal dispersion between batches, while maintaining minimal within batch dispersion. The procedure has been tested on a number of available datasets with and without prior information about batch partitioning. Datasets with a priori identified batches have been split accordingly, measured with weighted average Dice Index. Batch effect correction is justified by higher intra-group correlation. In the blank datasets, identified batch divisions lead to improvement of parameters and quality of biological information, shown by literature study and Information Content. The outcome of the algorithm serves as a starting point for correction methods. It has been demonstrated that omitting the essential step of batch effect control may lead to waste of valuable potential discoveries.

**Availability and implementation:** The implementation is available within the BatchI R package at http://zaed.aei.polsl.pl/index.php/pl/111-software.

**Supplementary information:** Supplementary data are available at Bioinformatics online.