**Bioinformatic approaches reveal the impact of major factors of the environmental metagenome extraction protocol**

Ji-Min Park1, Sung-Min Won1, In-Gyu Kim1, Jung-Hoon Yoon1*\**

*1 Department of Food Science and Biotechnology, Sungkyunkwan University, Jangan-gu, Suwon, Republic of Korea*

*\*Corresponding author: Jung-Hoon Yoon, jhyoon69@skku.edu*

**Highlights**

* Protocol on the metagenome extraction for various types of soil was standardized.
* Type and pH of buffer and reaction time were the most statistically significant.
* Quality of DNA extracted from new soil can be calculated from the formula.

**1. Introduction**

As is well known, the total environmental DNA (=Metagenome) is a key material in the biotechnology field with high added Value. Although many related researches were performed, much remains unclear about the factors affecting the quality of the extracted metagenome. Therefore, optimal methods or major factors for the direct isolation of high-quality DNA from various environments should be developed. So we introduce elaborate statistical techniques and use various bioinformatics tools to improve the extraction protocol and visualize the results. Seven variables from the adjustable parameters present in the protocol were selected, and the most statistically significant factors were the type of buffer, the pH of the buffer, the reaction time and the rate of centrifugation. The process design space and simulation about major factors and profiling was visualized. Our results provide a foundation for future research into clarifying the biodiversity in a specific soil microbial ecosystem, identification and mechanism of active genes and metabolites derived from metagenomes.

**2. Methods**

Considering some metagenomic DNA extraction protocols of recent related studies, we selected 14 various factors related to chemical reaction, physical force and soil sample. For experimental design, We first used the factorial-complete randomize design, factorial fractional design and taguchi method to check the main effects of some variables on the quality of the extracted DNA and their interaction effects with the samples. Since these statistical models needed some modifications, we customized the analysis design featuring the diversification of attributes and application of quality by design (QbD) and design of experiments (DOE). In data processing, missing values occurred when data collection having some bad influence on the results were dealt with by the 'Multiple imputation' technique. And then we carried out the simple correlation test, cluster analysis, cross tabulation analysis, Logistic regression analysis, conjoint analysis and principle components analysis to get the significant variables and their attributes related to the high quality metagenome. Finally, the simulation, profiling and process design space were performed to find the range of conditions for new soil sample in the future.

**3. Results and discussion**

**Table 1.** Summary of the cross tabulation, multiple regression and the cluster analysis related to the optimal group.



The analysis of the size (S), concentration (C), and both (T) of the extracted DNA was conducted. Although the some variable preferred different range of attributes by type of soil, the most of the variables had similar range of attributes. Within the experimental range, when the pH of the buffer is used with 8.0 for 7 hour and the rate of centrifugation was at 6,000rpm, generally high-quality metagenomic DNA was extracted.



**Figure. 1.** Optimization and simulation of variables with significant probability (a) The results of the prediction profiler, optimization; (b) The results of the surface profiler: ‘conc’ variable with pH and SDS, ‘size’ variable with pH and RT

When the forest soil sample is used, it was predicted that the condition to maximize both the concentration and size of the extracted DNA was to use a buffer (pH 5.8), SDS (3.6%) for 9 hour with no drying time.

**4. Conclusions**

The metagenome extraction protocol that can be applied to various environments was established through the statistical process based on six selected variables and their attributes. Using three representative types of sample, we present a range of major factors that can be adjusted to extract high-quality metagenomes from new soil samples. The visualization of adjustable range of major variables through the simulation and profiling makes inference of immediate results possible. And the efficiency of extracting high-quality metagenomes from various environments, creating libraries and exploring metagenomes is three times better than those of previous protocol.

**References**

1. Bertrand, H., Poly, F., Van, V.T., Lombard, N., Nalin, R., Vogel, T.M., and Simonet, P. (2005). High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. J Microbiol Meth 62, 1-11.
2. Lee, M.H., and Lee, S.W. (2013). Bioprospecting potential of the soil metagenome: novel enzymes and bioactivities. Genomics Inform 11, 114-120.
3. Nikolski, H.S.M. (2016). Machine learning for metagenomics methods and tools.pdf. 1-19.