УДК: 574.1:631.46:577.34

ПІДХОДИ ДО IN SILICO АНАЛІЗУ МЕТРИК РІЗНОМАНІТТЯ МІКРОБІОМУ ЗАБРУДНЕНИХ РАДІОНУКЛІДАМИ ҐРУНТІВ

О. Ю. ПАРЕНЮК[1], *кандидат біологічних наук*
І. О. СІМУТІН[2], Д. О. САМОФАЛОВА[3], Ю. В. РУБАН[1],
В. В. ІЛЛЄНКО[1], *кандидат біологічних наук*
Н. Г. НЕСТЕРОВА[1], *кандидат сільськогосподарських наук*
І. М. ГУДКОВ[1], *доктор біологічних наук*
[1]Національний університет біоресурсів і природокористування України
[2]Київський національний університет імені Тараса Шевченка, Київ
[3] ДУ «Інститут харчової біотехнології та геноміки НАН України»
E-mail: olena.pareniuk@gmail.com

Проаналізовано біоінформатичні підходи до обробки результатів секвенування тотальної ДНК ґрунтових мікроорганізмів. Підібрано методику, що забезпечує найвищу якість і достовірність вихідних даних.

Ключові слова: *мікробіом ґрунту, радіонуклідне забруднення, біоінформатичні методи*

Актуальність. Вивчення динаміки радіоактивних речовин у навколишньому середовищі та дослідження формування угруповань в умовах радіонуклідного забруднення є одним із завдань сучасної радіоекології, виконання якого дозволить впливати, зокрема, на міграцію радіонуклідів у довкіллі. Завдяки надзвичайно широкому видовому складу, швидкості зміни поколінь, а, отже, і високій швидкості еволюційних процесів, мікроорганізми ґрунту є однією з найменш досліджених, хоча і однією з найважливіших, ланок біосфери планети. Можливості прогнозування впливу мікроорганізмів на кислотність ґрунтів, вміст і форми органічної речовини, безпосереднє підвищення або пониження біологічної доступності певних сполук, зокрема радіонуклідів, визначають актуальність дослідження та побудови бази даних, що вміщуватиме відомості про наявні у ґрунті бактерії, їх належність до певних функціональних груп і місце кожної з таких груп у завершальній (редукуючій) ланці біологічного колообігу [1].

Останні досягнення у галузі геноміки та метагеноміки уможливлюють вивчення складних функцій мікробних спільнот в екосистемі з безпрецедентною точністю [2]. Деталізація даних, отриманих від NGS-аналізів, дає можливість вивчати структуру угруповання на різних філогенетичних рівнях – від царства до виду включно, залежно від якості виділених фрагментів ДНК і довжини отриманих рідів [3]. Це дає змогу вивчити зміни в екосистемах, спричинені варіюванням екологічних чинників на екосистемному рівні [4] і більш точно прогнозувати наслідки антропогенного впливу – наприклад, ефекти, спричинені глобальним потеплінням [4], забруд-

ненням важкими металами тощо [5]. Отже, це зручний інструмент для оцінки впливу радіоактивного забруднення на функції ґрунтової мікрофлори територій, відчужених від господарського використання внаслідок аварій, пов'язаних з надходженням радіоактивних речовин у навколишнє середовище.

Методика обробки отриманих при секвенуванні біоінформатичних даних суттєво впливає на якість отриманих даних. Так, неправильний аналіз може спричинити невірну інтерпретацію і, отже, низьку достовірність отриманих даних.

Дана робота пропонує методику аналізу біоінформатичних даних, що була протестована на зразках, відібраних з саркофагу над зруйнованим 4-м енергоблоком ЧАЕС. У статті запропоновано методи вирішення наступних задач:

1) аналіз якості даних секвенування;
2) проведення препроцесінгу даних секвенування;
3) проведення кластеризації метагеномних даних;
4) розрахунок альфа різноманіття мікробіому;
5) розрахунок бета-різноманіття мікробіому;
6) проведення функціональної реконструкції мікробіому;
7) підбір методів візуалізації метагеномних даних.

Запропонована послідовність дій буда протестована та дозволила отримати валідні та достовірні дані.

Матеріали і методи. У жовтні 2015 року було відібрано 8 зразків субстрату на стандартних маршрутах моніторингу Інституту проблем безпеки АЕС НАН України в приміщеннях зруйнованого 4 енергоблоку ЧАЕС. Зразки 2, 5 та 8 були відібрані з приміщення під шахтою реактора, 4 та 12 – за межами об'єкту «Укриття» і використовувалися як контроль, 11 та 7 – із внутрішніх виходів шахт системи моніторингу реактора «Фініш», 13 – із дренажної труби, що виходить з об'єкту «Укриття».

У результаті секвенування описаних зразків по амплікону 16S рРНК були отримані 12 файлів за прямими і зворотними рідами з використанням платформи Illumina MiSeq. Отримані дані секвенування ДНК були представлені у FastQ форматі, для аналізу використовувалися зразки тільки прямих прочитань - *R1 (Табл. 1).

Для аналізу був використаний комп'ютер наведеної нижче конфігурації:

1. Процесор – 2.5 GHz Intel Core I7.
2. Пам'ять – 16 GB 1600 MHz DDR3.
3. Графіка – Intel Iris Pro 1536 MB.

Версія операційної системи: MAC OS Sierra V10.12

Статична обробка даних була здійснена за допомогою програмних пакетів FastQC та QIIME.

1. Технічна характеристика досліджуваних зразків

Назва файлу	Загальна кількість нуклеотидів, п.о.	Вміст ГЦ, %	Довжина сиквенсу	Розмір файла, Мб
4365-R4-MS357wF_R1.fastq	39669	58	36-251	21,5
4365-R5-MS357wF_R1.fastq	53364	57	35-251	26,7
4365-R7-MS357wF_R1.fastq	32708	56	37-251	18,3
4365-R11-MS357wF_R1.fastq	29135	55	37-251	18,3
4365-R12-MS357wF_R1.fastq	11914	57	165-251	6,7
4365-R13-MS357wF_R1.fastq	33404	57	38-251	18,8

Результати дослідження та їх обговорення. Загальна схема аналізів. Базуючись на загальноприйнятій методології метагеномних досліджень (рис. 1), першочергово необхідно перевірити якість даних результатів секвенування за допомогою програмного пакету FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) [6]. Більшість секвенаторів генерують так званий QC (quality control) звіт як необхідну частину її схеми аналізу, але цей звіт, як правило, сфокусовано на ідентифікації проблем, які генеруються самим секвенатором. Програмний пакет FastQC забезпечує QC звіт, що дає можливість визначити проблеми, які генеруються безпосередньо секвенатором або матеріалами стартової бібліотеки. Наступним етапом є демультиплікація даних секвенування за допомогою біоінформатичного програмного пакету QIIME (Quantitative Insights Into Microbial Ecology) [7] модулем *split_libraries_fastq.py*.

QIIME – це набір функціонально різних модулів із відкритим кодом, що написані за допомогою мови програмування Python та спеціально розроблені для аналізу метагеномних даних секвенування. Модулі QIIME виконуються тільки з командного рядка та не мають графічного інтерфейсу. Головними можливостями QIIME є демультиплекс та фільтрування якості даних (demultiplexing, quality filtering), кластерізація (OTU picking), таксономічне розподілення (taxonomic assigmnent), філогенетична реконструкція (phylogenetic reconstruction), аналіз

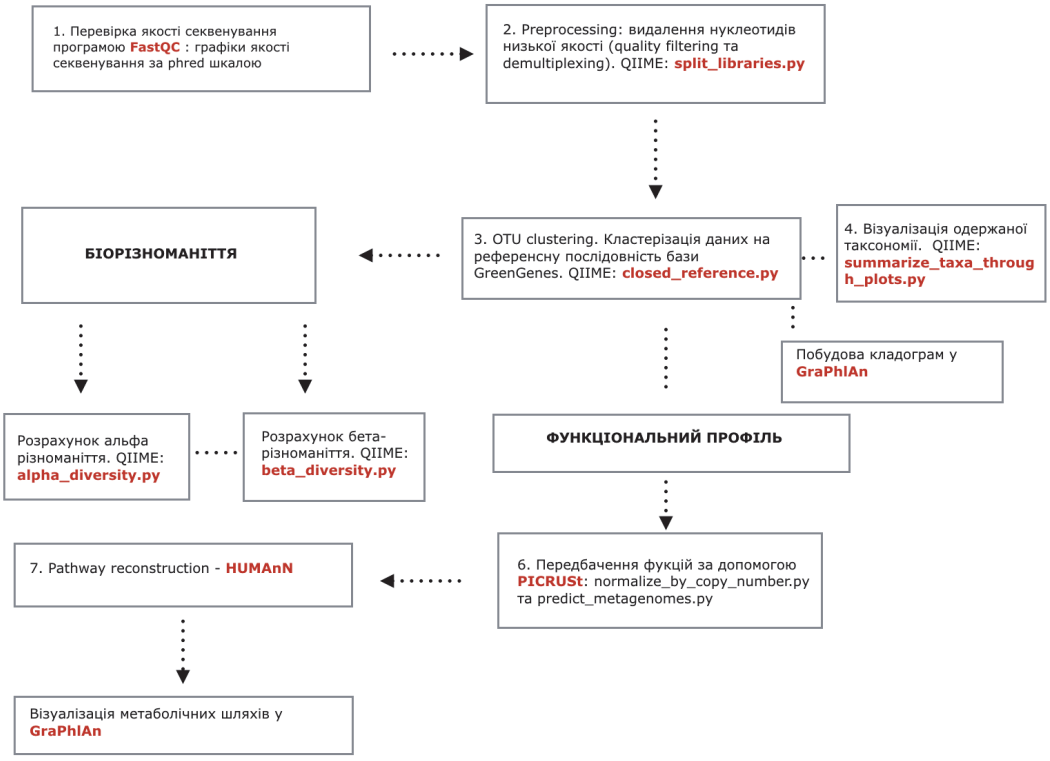

Рис. 1. Загальна схема аналізу.

різноманіття та візуалізація. Далі, відфільтровані сіквенси за якістю кластеризуються на референсну послідовність (в даному випадку база даних 16s РНК послідовностей – Greengenes) модулем closed_reference.py. Одержані OTU таблиці візуалізуються за допомогою модуля summarize_taxa_ though_plots.py. Додатково будуються кладограми у програмі GraPhlAn [8].

Аналіз біорізноманіття розраховується модулями QIIME – відповідно alpha_diversity.py для альфа-різноманіття та beta_diversity.py для бета-різноманіття.

Наступним кроком є передбачення біологічних функцій та реконструкція метаболічних шляхів у програмах PICRUSt [9] та HUMANn. Одержані дані метаболічних шляхів візуалізуються за допомогою програми GraPhlAn.

Перевірка якості даних секвенування програмою FastQC. Програма FastQC генерує звіти щодо якості секвенування у вигляді графіків при завантаженні даних та обранні опції "save report".

Висновки щодо якості базуються на аналізі певних параметрів діапазону значення якості всіх основ у кожній позиції FastQ-файлу (Per Base Sequence Quality), що є визначальним графіком контролю якості та діаграмою розрахунку. Елементами цього графіку є: центральна червона лінія – медіана, жовті блоки – діапазон між квартилами, блакитна лінія – середня якість, вуса – мінімум та максимум, що спостерігаються. За віссю абсцис відкладається позиція нуклеотиду у *.fastq файлі, за віссю ординат – значення якості, що вимірюється у шкалі Phred [10]. Додаткові параметри якості, що є другорядними та розшифровують значення головного графіку:

1) універсальне негативне значення якості (Per Sequence Quality Scores);

2) показник пропорції для кожного нуклеотиду та його позиції у файлі (Per Base Sequence Content);

3) кількість гуаніну та цитозину вздовж всієї довжини для кожного сиквенсу (Per Sequence GC Content);

4) показник рівня дуплікації (Duplicate Sequences).

Препроцесінг метагеномних даних у програмному пакеті QIIME. Препроцесінг даних [10] – це етап підготовки даних секвенування до кластеризації та подальших етапів аналізу. А саме: quality filtering – видалення довгих нехарактерних послідовностей (sequence lengths), енд-тріммінг (end-trimming) та видалення всіх послідовностей, що відповідають низькій якості за шкалою phred (minimum quality score). У даному випадку був застосований модуль QIIME split_libraries_fastq.py (http://qiime.org/scripts/split_libraries_fastq.html), що потребує вхідних даних у форматі *.fastq та має певні особливості щодо вибору фільтрів ((p), r, (n), (q), c).

Для кожного *.fastq файлу виконується препроцесінг наступною командою:

*(qiime1)./split_libraries_fastq.py -i *.fastq –sample_ids R* –min_per_read_length_fraction 0.75 –phred_quality_threshold 19 –barcode_type "not-barcoded" -o (output directory)*

Відповідно обираються параметри min_per_read_length_fraction зі значенням 0,75 та phred_quality_threshold зі значенням 19. Додатково вибирається параметр barcode_type «not-barcoded», тому що баркоди (короткі послідовності необхідні для ідентифікації під час секвенування [11]) видаляються на попередніх етапах.

Кластеризація метагеномних даних у QIIME. Метою даного етапу є отримання OTU (operational taxonomic unit) таблиць, що містять кластеризовану інформацію щодо бактеріальної таксономії базуючись на подібності послідовностей (sequence

similarity) [12]. У 16S метагеномних підходах OTU є кластером варіантів подібних послідовностей 16S rDNA маркерних генів. Для генерування OTU таблиць використовується модуль QIIME closed_reference.py (http://qiime.org/scripts/pick_closed_reference_otus.html). Особливістю даного етапу є те, що вирівнювання відбувається за базою даних послідовностей 16S "Greengenes" (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) за замовчуванням, але є можливість обрати іншу базу (наприклад, "SILVA"(https://www.arb-silva.de/)).

Для кожного *.fastq файлу виконується кластеризація наступною командою:

Як результат генеруються OTUs таблиці у *.biom форматі [12] – The Biological Observation Matrix (http://biom-format.org/).

(qiime1)./pick_closed_reference_otus.py -i seqs.fna -o (otput directory)

Розрахунок альфа-різноманіття мікробіому. Альфа-різноманіття дає можливість оцінити біорізноманіття (number of taxa) у межах однієї спільноти та розрахувати кількість детектованої таксономії у конкретному зразку. Альфа-різноманіття має низку можливих метрик для оцінки різноманіття, а саме: chao1, shannon, simpson, dominance тощо (http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html).

Для розрахунку альфа-різноманіття застосовується модуль QIIME alpha_diversity.py (http://qiime.org/scripts/alpha_diversity.html), що потребує на вході *.biom таблиці та додаткову інформацію щодо кладистики в форматі *.tre. Для кожного зразка альфа-різноманіття розраховується наступною командою:

(qiime1)./alpha_diversity.py -i otu_table.biom -m chao1,PD_whole_tree, shannon, simpson,dominance -o (output directory)

Як видно з команди, для розрахунку обираються наступні метричні параметри(m): chao1, PD_whole_tree, shannon, simpson, dominance.

Графіки для альфа різноманіття будуються у пакеті R phyloseq [13] функцією plot_richness (otu,x=»samples»,measures=c («Chao1»,»Shannon»,»Simpson»)).

Розрахунок бета-різноманіття мікробіому. Бета-різноманіття дозволяє оцінити різноманіття між декількома спільнотами. Як і альфа-різноманіття, так і бета-різноманіття має свої специфічні метричні параметри: bray_curtis, chisq, euclidean, kulczynski, binary_sorensen_dice та інші.

Для розрахунку бета-різноманіття застосовується модуль QIIME beta_diversity.py (http://qiime.org/scripts/beta_diversity.html), що потребує на вході *.biom таблиці та вказані метричні параметри. Для розрахунку бета-різноманіття застосовується злита *.biom таблиця (merged), що містить *.biom таблиці всіх зразків. Розрахунок здійснюється наступною командою:

(qiime1)./beta_diversity.py -i merged.biom -m bray_curtis, chisq, euclidean, kulczynski, binary_sorensen_dice -o output directory

Як видно з команди, для аналізу обираються наступні метричні параметри: bray_curtis, chisq, euclidean, kulczynski, binary_sorensen_dice.

Візуалізація бета-різноманіття робиться за допомогою PCoA графіків за допомогою модуля QIIME principal_coordinates.py (http://qiime.org/scripts/principal_coordinates.html) та make_2d_plots.py (http://qiime.org/scripts/make_2d_plots.html), та виконується наступними командами:

Висновки. Отримані дані секвенування мікробіому зруйнованого 4 енергоблоку ЧАЕС були проаналізовані за якістю секвенування програмним пакетом

FastQC. Встановлено, що якість даних секвенування коливається у межах 10-48 за phred шкалою. Проведено препроцесінг даних секвенування у програмі QIIME та видалено послідовності, показник якості яких нижчий за 19 одиниць за phred шкалою. Проведено кластеризацію метагеномних даних на основі рефенсну у програмі QIIME та отримано OTU таблиці і результати кладистики для досліджуваних зразків. Підібрані методи аналізу метагеномних даних дозволяють отримувати достовірний опис складу та функцій мікробіому.

Література

1. Pareniuk O. Modification of 137cs transfer to rape (brassica napus l.) phytomass under the influence of soil microorganisms / O. Pareniuk, K. Shavanova, J. P. Laceby[et al.] // Journal of Environmental Radioactivity. — 2015. — Vol. 149. — P. 73–80.
2. Schneider T. Who is who in litter decomposition? metaproteomics reveals major microbial players and their biogeochemical functions. / T. Schneider, K. M. Keiblinger, E. Schmid[et al.] // The ISME journal. — 2012. — Vol. 6, No. 9. — P. 1749–62.
3. Thomas T. Metagenomics - a guide from sampling to data analysis. / T. Thomas, J. Gilbert, F. Meyer // Microbial informatics and experimentation. — 2012. — Vol. 2, No. 1. — P. 3.
4. Luo C. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics.1. luo c. soil microbial community responses to a decade of warming as revealed by comparative metagenomics. / c. luo, l. m. rodriguez-r, e. r. johnston[et a / C. Luo, L. M. Rodriguez-R, E. R. Johnston[et al.] // Applied and environmental microbiology. — 2014. — Vol. 80, No. 5. — P. 1777–86.
5. Gołębiewski M. 16s rdna pyrosequencing analysis of bacterial community in heavy metals polluted soils. / M. Gołębiewski, E. Deja-Sikora, M. Cichosz[et al.] // Microbial ecology. — 2014. — Vol. 67, No. 3. — P. 635–47.
6. Schmieder R. Quality control and preprocessing of metagenomic datasets / R. Schmieder, R. Edwards // Bioinformatics. — 2011. — Vol. 27, No. 6. — P. 863–864.
7. Kuczynski J. Using qiime to analyze 16s rrna gene sequences from microbial communities. / J. Kuczynski, J. Stombaugh, W. A. Walters[et al.] // Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]. — 2011. — Vol. Chapter 10. — P. Unit 1E.5.
8. Asnicar F. Compact graphical representation of phylogenetic data and metadata with graphlan / F. Asnicar, G. Weingart, T. L. Tickle[et al.] // PeerJ. — 2015. — Vol. 3. — P. e1029.
9. Langille M. G. I. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences / M. G. I. Langille, J. Zaneveld, J. G. Caporaso[et al.] // Nature Biotechnology. — 2013. — Vol. 31, No. 9. — P. 814–821.
10. Ewing B. Base-calling of automated sequencer traces using phred. ii. error probabilities. / B. Ewing, P. Green // Genome research. — 1998. — Vol. 8, No. 3. — P. 186–94.
11. Puente-Sánchez F. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. / F. Puente-Sánchez, J. Aguirre, V. Parro // Nucleic acids research. — 2016. — Vol. 44, No. 4. — P. e40.
12. Blaxter M. Defining operational taxonomic units using dna barcode data / M. Blaxter, J. Mann, T. Chapman[et al.] // Philosophical Transactions of the Royal Society B: Biological Sciences. — 2005. — Vol. 360, No. 1462. — P. 1935–1943.
13. McMurdie P. J. Phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data / P. J. McMurdie, S. Holmes, R. Kindt[et al.] // PLoS ONE. — 2013. — Vol. 8, No. 4. — P. e61217.

References

1. Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., & Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ, 3, e1029. http://doi.org/10.7717/peerj.1029
2. Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the

Royal Society B: Biological Sciences, 360(1462), 1935–1943. http://doi.org/10.1098/rstb.2005.1725
3. Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research, 8(3), 186–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9521922
4. Gołębiewski, M., Deja-Sikora, E., Cichosz, M., Tretyn, A., & Wróbel, B. (2014). 16S rDNA pyrosequencing analysis of bacterial community in heavy metals polluted soils. Microbial Ecology, 67(3), 635–47. http://doi.org/10.1007/s00248-013-0344-7
5. Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., & Knight, R. (2011). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.], Chapter 10, Unit 1E.5. http://doi.org/10.1002/9780471729259.mc01e05s27
6. Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., ... Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology, 31(9), 814–821. http://doi.org/10.1038/nbt.2676
7. Luo, C., Rodriguez-R, L. M., Johnston, E. R., Wu, L., Cheng, L., Xue, K., ... Konstantinidis, K. T. (2014). Soil microbial community responses to a decade of warming as revealed by comparative metagenomics.1. Luo C. Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. / C. Luo, L. M. Rodriguez-R, E. R. Johnston[et al.] Applied and Environmental Microbiology, 80(5), 1777–86. http://doi.org/10.1128/AEM.03712-13
8. McMurdie, P. J., Holmes, S., Kindt, R., Legendre, P., & O'Hara, R. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE, 8(4), e61217. http://doi.org/10.1371/journal.pone.0061217
9. Pareniuk, O., Shavanova, K., Laceby, J. P., Illienko, V., Tytova, L., Levchuk, S., ... Nanba, K. (2015). Modification of 137Cs transfer to rape (Brassica napus L.) phytomass under the influence of soil microorganisms. Journal of Environmental Radioactivity, 149, 73–80. http://doi.org/10.1016/j.jenvrad.2015.07.003
10. Puente-Sánchez, F., Aguirre, J., & Parro, V. (2016). A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. Nucleic Acids Research, 44(4), e40. http://doi.org/10.1093/nar/gkv1113
11. Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27(6), 863–864. http://doi.org/10.1093/bioinformatics/btr026
12. Schneider, T., Keiblinger, K. M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., ... Riedel, K. (2012). Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. The ISME Journal, 6(9), 1749–62. http://doi.org/10.1038/ismej.2012.11
13. Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation, 2(1), 3. http://doi.org/10.1186/2042-5783-2-3

SUMMARY

O. Parenyuk, I. Simutin, D. Samofalova, Yu. Ruban, V. Illienko, N. Nesterova, I. Gudkov. *Approaches to in silico analysis of micobiome biodiversity metrics of radionuclide contaminated soils/ Biological Resources and Nature Managment.– 2017. – **9**, №5–6. – P.10–16.*

The bioinformatic approaches to the processing of the results of total soil microorganisms DNA sequencing have been analyzed. The methodology that provides the highest quality and reliability of the source data was designed.

Keywords: *microbiology of soil, radioactive contamination, bioinformatic methods*

АННОТАЦИЯ

Е. Ю. Паренюк, И. О. Симутин, Д. А. Самофалова, Ю. В. Рубан, В. В. Илльенко, Н. Г. Нестерова, И. Н. Гудков. *Подходы к in silico анализу метрик разнообразия микробиома загрязненных радионуклидами почв//Биоресурсы и природопользование. – 2017. – **9**, №5–6. – С.10–16.*

Проанализированы биоинформатические подходы к обработке результатов секвенирования тотальной ДНК почвенных микроорганизмов. Подобрана методика, которая обеспечивает наивысшее качество и достоверность исходных данных.

Ключевые слова: *микробиом почвы, радиоактивное загрязнение, биоинформатические методы*