Functional variation among bacterial taxa within host-associated microbiomes

Supervisor: Aaron Comeault

Tobias Nunn

## Aims

Aim 1: determine whether bacterial genera living on or in different hosts have different genomic functional capacities.

Aim 2: determine whether different bacterial taxa (genera or families) differ in their predicted functional capacity independent of their host environment.

## Abstract

Bacterial Microbiomes are cryptic environments, in which bacteria exhibit cooperative and antagonistic interactions. Understanding these relationships could have consequences for many other fields, including health and conservation, due to the universal presence of bacteria and the functional roles they play in their environments. Due to the microscopic nature of these ecosystems, genomics is a useful way to study them, and much progress has been made on this in the field’s 30 year history.

Publicly available genomic databases hold vast quantities of information that can assist in drawing new conclusions about the functional differences between taxa, specifically species and genus, and how their environment may shape these differences, and how this may impact potential hosts. This study shall find these differences by annotating many public bacterial genomes and using KEGG values to find functional differences. This will lead to a greater understanding of the functional roles bacteria play in these environments.

## Introduction

Bacterial genomics is an interesting field and there has been much evolution over the 30 years the field has existed. The first bacterial genome to be sequenced was that of *H. influenzae* in 1995 using shotgun sequencing (Fleischmann et al., 1995). Since then there have been many technological improvements, allowing genomes to be sampled from over 50 bacterial phyla, with continually more data being reported about lesser researched taxonomic groups (Land et al., 2015). Much research has been done into the relationships between bacteria and host. For example, nutrient acquisition, the digestion of complex compounds, such as plant polysaccharides and the production of amino acids and vitamins, useful for the host. Often, these are functions that hosts cannot perform for themselves (Mendes and Raaijmakers, 2015). A greater understanding of this can have implications for many fields, such as: bioforensics; disease treatment; biosurveillance and metabolic modeling (Land et al., 2015). There have been continually new generations of sequencing technology. The most modern and powerful is Oxford Nanopore (Abdi et al., 2024). Many studies have utilised these methods, however few discuss microbial genomics in the context of relationships and interactions between microbiomes and their environment or host, and the factors that affect these relationships. A study on the microbiome of a species of potato plant (*Solanum tuberosum*) discussed how certain species of bacteria can alleviate the impacts of drought on crop yield. Potatoes are an important food globally and ensuring food stability is vital for future development (Petrushin et al., 2024). As climate change worsens, water scarcity will become more common in some regions, so an understanding of how to make them more plastic to this effect is important.

The human gut microbiome has been studied frequently, probably due to the relationship between it and public health. For example Crohn’s disease is significantly linked to lower gut microbiome richness. Similarly, a relationship between microbial load and variation was found (Vandeputte et al., 2017). An understanding of Microbiomes can have implications for conservation efforts. For example, in combatting Chytridiomycosis, an infectious disease causing amphibian population decline. Previous research found that differences in skin microbiome can impact on susceptibility to infection (Jiménez and Sommer, 2017). This is supported by Harris et al. (2009), who also found that the skin microbiome is an important factor in disease survival. There are other important elements to the skin of amphibians, for example respiration, that the microbiome affects, making a greater understanding of the relationships and functions therein vital (Harris et al., 2009). Another study found that members of the same species showed similar microbiomes, whereas there were fewer similarities between different individuals from the same pond (McKenzie et al., 2012). This might be evidence of co-evolution between host species and microbiome. The practical uses of bacteria in the context of sustainability are varied. A recent study outlined how bacteria are being used to recycle waste gases or biomass and produce products like ethanol, other researchers have found bacteria that break down certain plastics (Rappuoli et al., 2025).

Relationships between bacteria have been studied previously, in the context of cooperation in biofilm formation. Both antagonism and cooperation were found in the bacterial species studied. The sample was taken from drinking water, six species were identified and 96 agar plates with different combinations of these were created. Previous studies had found that cooperation lead to more stable communities and a few specific beneficial relationships were identified, for example, plasmid conjugation (Simões et al., 2007). A recent study by (Lunjani et al., 2021) discussed host-microbe interactions in human skin with relevance to how modern factors are changing these relationships and causing inflammatory skin diseases. The study outlines how this knowledge could help in creating personalised probiotic solutions to this problem. There are many ways in which bacteria in a microbiome can interact, as mentioned previously, they can cooperate and produce biofilms and share nutrients, however, there also exists competition wherever there is a niche to exploit and limited resources. These relationships cause selection for genes that cause, for example, the production of antimicrobial compounds to inhibit the competitors.

There are many species of bacteria, as such there are almost limitless ways in which they interact with each other in different conditions, gaining better insight into some of the ways in which these relationships occur on a genetic level could have impacts for fields as broad as public health to pollution to zoonotic disease and conservation. This study will aim to investigate some of these relationships through bottom-up estimates of function based on gene content, using publically available data, and present these in the context of how they might have a beneficial impact on potential host organisms. This work may be cause for future research to be done in the area.

## Methods

This project is desk-based, and all data analysis will be undertaken computationally. Thus, data must be collected from online sources. The online genome data repository used by this project is the National Center for Biotechnology Information (NCBI) [genome database](https://www.ncbi.nlm.nih.gov/datasets/genome/). There are roughly 2.61 million bacterial genomes in this database that could be annotated. Along with this exists a range of metadata from taxonomy information to host and location data. This provides a benefit as the time to pull down this data is much shorter than the time taken to harvest samples from the field, then grow and sequence them in the lab. With this I can create novel analyses and draw new conclusions about how the evolutionary history shapes the relationships between bacteria. Another important database is the Kyoto Encyclopedia of Genes and Genomes (KEGG) [database](https://www.genome.jp/kegg/). This is important as it holds data on the function of genes, these are grouped into “Orthologs” ([KO values](https://www.genome.jp/kegg/ko.html)) of shared function. The database also has data on how these KOs fit into larger biological systems, such as metabolism. These are called “[pathways](https://www.genome.jp/kegg/pathway.html)”. With this information, i can directly correlate the genes possessed by different species of bacteria to their potential functions in the bacteria and thus, draw conclusions about how those functions could shape relationships with potential hosts and other bacteria in the environment. Specifically, grouped analysis comparing genera will be done to visualise how the evolutionary differences between genera can cause differential genome content and gene function.

In order to compare the genomes against the KEGG database, they must first be transformed. Firstly, [EggNOG-mapper/2.1.12](https://github.com/eggnogdb/eggnog-mapper/blob/master/README.md) will be utilised to annotate the genomes using slurm scripting, this will identify the genes of the genome. After this, the R package [MicrobiomeProfiler](https://bioconductor.org/packages/release/bioc/html/MicrobiomeProfiler.html) will be used, specifically the EnrichKO() function, which will [enrich](https://rdrr.io/github/YuLab-SMU/MicrobiomeProfiler/man/enrichKO.html) the annotated genomes so as to identify the significant KEGG pathways, their functions can be discovered through comparison with the database. From this point, visualisations such as heatmaps will be used to present these pathways. In order to accomplish this, groupings will need to be created, such as genera inside a specific family of bacteria or host taxonomy. These can then be analysed in the context of the evolutionary history or environment of the groups selected. Using this wealth of information. A secondary item of this project is to explore alternatives, microbial genomics is a cutting edge field, with more tools emerging continuously, and finding which systems or processes work best for these analysis will be important for continued research. For example, the enrichment tool [KofamKOALA](https://www.genome.jp/tools/kofamkoala/) or the [gene ontology (GO)](https://geneontology.org/) system have been identified as possible alternatives, and can be compared with the current methods to see which provides the best data clarity, accuracy and rapidity.

Due to the size of the dataset, efficient storage of this data will be vital so that the computational systems do not get overwhelmed. To aid this, a relational database will be created in PostgreSQL. PostgreSQL is good at handling the .JSON data format which is used in many public APIs, including the NCBI. A relational database will allow questions about the data to be answered by querying; host information, taxonomy, etc. For the purposes of data transparency and experiment reproducibility, the scripts used in this project will be uploaded to a github repo for the project. So that any researchers who wish to build on this project can take inspiration from or adapt the scripts used. This repo will also contain a series of notebooks outlining where the data came from, the specifics of the data pipeline and the significance of the outputs. API calls are efficient ways of transferring large sums of information, which will allow for more complex analysis to be performed. Due to the lack of manual steps in this, the code should provide clarity for future researchers to more easily reproduce or understand the step.

Finally, the data will be supplemented by over 100 genomes generated in the labs at ECW by Bangor Masters and PhD researchers. These samples come from amphibian skin microbiomes. The known status of these samples mean that more exact analysis can be performed, as well as providing good comparisons. For example, there are not many samples from amphibian skin microbiomes in the database. Thus, a comparison of how the same genera differ when on these hosts compared with the environment could be created. This will lead to other types of visualisation to be created, for example, phylogenetic trees to confirm the taxonomic assignment of these samples.

## References

Abdi, G., Tarighat, M.A., Jain, M., Tendulkar, R., Tendulkar, M. and Barwant, M. 2024. Revolutionizing Genomics: Exploring the Potential of Next-Generation Sequencing *In*: V. Singh and A. Kumar, eds. *Advances in Bioinformatics* [Online]. Singapore: Springer Nature, pp.1–33. [Accessed 13 March 2025]. Available from: <https://doi.org/10.1007/978-981-99-8401-5_1>.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O. and Venter, J.C. 1995. Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd *In*: *Science* [Online]., pp.496–512. [Accessed 13 March 2025]. Available from: <https://www.science.org/doi/abs/10.1126/science.7542800>.

Harris, R.N., Brucker, R.M., Walke, J.B., Becker, M.H., Schwantes, C.R., Flaherty, D.C., Lam, B.A., Woodhams, D.C., Briggs, C.J., Vredenburg, V.T. and Minbiole, K.P.C. 2009. Skin microbes on frogs prevent morbidity and mortality caused by a lethal skin fungus *In*: *The ISME Journal* [Online]., pp.818–824. [Accessed 10 March 2025]. Available from: <https://doi.org/10.1038/ismej.2009.27>.

Jiménez, R.R. and Sommer, S. 2017. The amphibian microbiome: Natural range of variation, pathogenic dysbiosis, and role in conservation *In*: *Biodiversity and Conservation* [Online]., pp.763–786. [Accessed 9 April 2025]. Available from: <https://doi.org/10.1007/s10531-016-1272-x>.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S. and Ussery, D.W. 2015. Insights from 20 years of bacterial genome sequencing *In*: *Functional & Integrative Genomics* [Online]., pp.141–161. [Accessed 13 March 2025]. Available from: <https://doi.org/10.1007/s10142-015-0433-4>.

Lunjani, N., Ahearn-Ford, S., Dube, F.S., Hlela, C. and O’Mahony, L. 2021. Mechanisms of microbe-immune system dialogue within the skin *In*: *Genes & Immunity* [Online]., pp.276–288. [Accessed 9 March 2025]. Available from: <https://www.nature.com/articles/s41435-021-00133-9>.

McKenzie, V.J., Bowers, R.M., Fierer, N., Knight, R. and Lauber, C.L. 2012. Co-habiting amphibian species harbor unique skin bacterial communities in wild populations *In*: *The ISME Journal* [Online]., pp.588–596. [Accessed 14 March 2025]. Available from: <https://doi.org/10.1038/ismej.2011.129>.

Mendes, R. and Raaijmakers, J.M. 2015. Cross-kingdom similarities in microbiome functions *In*: *The ISME Journal* [Online]., pp.1905–1907. [Accessed 7 April 2025]. Available from: <https://doi.org/10.1038/ismej.2015.7>.

Petrushin, I.S., Filinova, N.V. and Gutnik, D.I. 2024. Potato Microbiome: Relationship with Environmental Factors and Approaches for Microbiome Modulation *In*: *International Journal of Molecular Sciences* [Online]., p.750. [Accessed 28 March 2025]. Available from: <https://www.mdpi.com/1422-0067/25/2/750>.

Rappuoli, R., Nguyen, N.K., Bloom, D.E., Brooke, C., Burckhardt, R.M., Dangour, A.D., Egamberdieva, D., Gronvall, G.K., Lawley, T.D., Lennon, J.T., Morhard, R., Mukhopadhyay, A., Peixoto, R., Silver, P.A. and Stein, L.Y. 2025. Microbes can capture carbon and degrade plastic — why aren’t we using them more? *In*: *Nature* [Online]., pp.864–866. [Accessed 30 March 2025]. Available from: <https://www.nature.com/articles/d41586-025-00875-w>.

Simões, L.C., Simões, M. and Vieira, M.J. 2007. Biofilm Interactions between Distinct Bacterial Genera Isolated from Drinking Water *In*: *Applied and Environmental Microbiology* [Online]., pp.6192–6200. [Accessed 28 March 2025]. Available from: <https://journals.asm.org/doi/full/10.1128/aem.00837-07>.

Vandeputte, D., Kathagen, G., D’hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G. and Raes, J. 2017. Quantitative microbiome profiling links gut community variation to microbial load *In*: *Nature* [Online]., pp.507–511. [Accessed 28 March 2025]. Available from: <https://www.nature.com/articles/nature24460>.

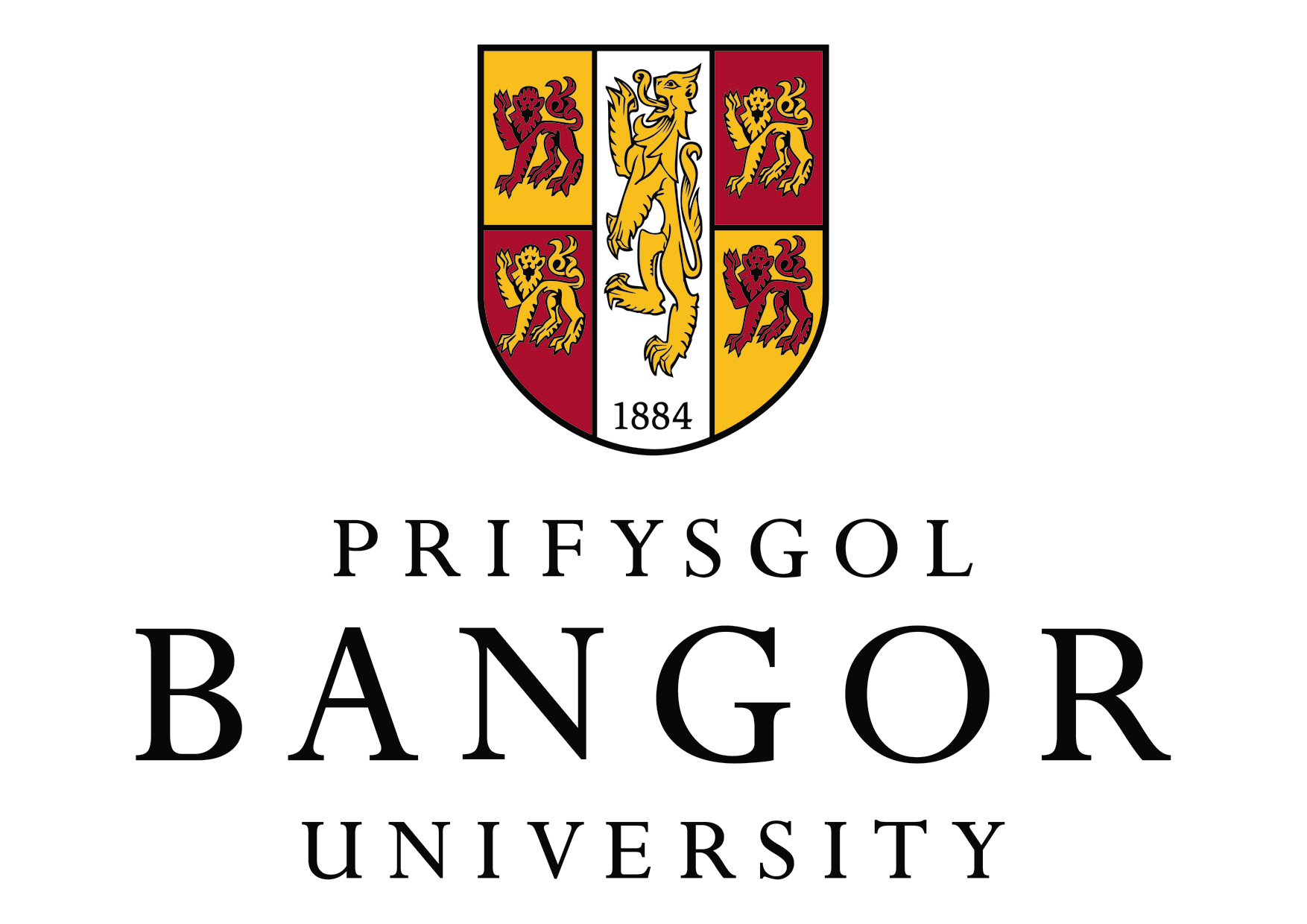
**A logo of a university

AI-generated content may be incorrect.**Risk Assessment Form

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Location / Building / Area:** | Accommodation | | **Activity (Summary):** | Using various computational methods to make novel discoveries about bacteria |
| **Delete as appropriate** | **Henfaes** No | **Treborth** No |
| **Date of Assessment:** | 2025-03-24 | |
| **Academic supervisor** | Aaron Comeault | | **Name of student:** | Tobias Nunn |
| **Signature:** | Aaron Comeault | | **Signature:** | TNUNN |

School of Natural Science (Environment)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **What are the**  **dangers/hazards?** | **Who might be harmed ?** | **What are you already doing to prevent harm?** | **What further action is necessary?** | **Action by whom** |
| Muscle strain  RSI | Self | * Ensure computer and chair correctly adjusted * Use docking station for extended laptop use | Take regular breaks | Self |
| Eye strain | Self | Ensure monitor is well lit, clear and free from flicker | Take regular breaks | Self |
| Electrocution | Self | Ensure all plugs and cables are safe, undamaged, no exposed wires | Check extension leads not overloaded | Self |
| stress | Self | Be mindful of how I am spending my time | Take regular breaks | Self |

**College of Science and Engineering**

**RESEARCH PROJECT ETHICAL ISSUES CHECKLIST FOR STAFF AND STUDENTS**

To assist with record keeping, please name your checklist files according to the following format: “**Ethics\_[surname]\_[year]”.**

It is your responsibility to follow the University’s Research Ethics Policy, the General Data Protection Regulation, and any relevant academic or professional guidelines in the conduct of your study[[1]](#footnote-1). It is also your responsibility to ensure that you have all necessary permits to conduct your research.

|  |  |  |
| --- | --- | --- |
| **A. Research involving people or biological samples from people** | **YES** | **NO** |
| Does the proposed research involve human participants, human biological samples, personal data or records accessed by interviews, lab experiments, interventions, surveys, observation or taken from social media etc*?*    *This includes any form of social research (interviews, questionnaires, observations, data dredging from social media etc.)*  *If yes, consult the Ethics Rep in your school or the CoSE Ethics Chair to see if your research is covered by a Grouped Activity before proceeding.* |  | No |
| **B. Research on animals / imported plants** | **YES** | **NO** |
| 1. Does the proposed research involve live vertebrates or cephalopods? |  | No |
| 1. Does the proposed research involve live invertebrates other than cephalopods? |  | No |
| 1. Does the proposed research involve the import of material (plants, animal, soil or sediments) from outside the UK? |  | No |

Checklist continues on next page…

|  |  |  |
| --- | --- | --- |
| C. Research using data or materials freely available in the public domain (e.g. newspaper and media reports, social media posts, published records from administrative sources) | **YES** | **NO** |
| 1. Are the materials covered by a clearly defined license or copyright exemption?  If in doubt, please consult the Ethics Co-Chair within your school or the CoSE Ethics Chair. | **YES** |  |
| 1. Does the proposed research using the freely available data involve sensitive topics such as protected personal characteristics (age, disability, gender reassignment, pregnancy and maternity, race, religion or belief, sex and sexual orientation), evidence of immoral, illegal or unethical behaviour, or investigating closed social groups (such as clubs, cliques, gangs, etc.)? |  | **No** |
| 1. Will the proposed research using the freely available data create a derivative work (as defined by the Copyright, Designs and Patents Act 1988 amended)?   A derivative work usually involves transforming the original data in some way. Simple observation (counting, reading, summarising) does not constitute a derivative work. Models trained on the data will usually create a derivative work. |  | **No** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Permission to conduct your research** | **YES** | **NO** | **N/A** |
| Please confirm you have all required permission to conduct your research (including permission from land owners to access land). Please note it is your responsibility to ensure all permits and permission is obtained. |  |  | N/A |

**If Sections A or B apply to your research:**

If you answered **‘NO’ in ALL of sections A AND B**, please keep this form on file as it may need to be referred to when you submit your thesis or results for publication. You **do not** need to submit an application through the online Research Ethics Application System.

If you answer **‘YES’ in any of sections A OR B,** you **do** need to submit an application through the online Research Ethics Application System. **First,** draft your ethics application using the offline template.

**If Section C applies to your research:**

If you select any of the red shaded boxes, please contact the Ethics Rep in your school or the CoSE Ethics Chair. You will need the following:

1. A link to the materials/data you propose to use.
2. A description of how the materials/data will be used.

You will then be advised on whether a Research Ethics Application will be required.

1. *For more, and the latest information, please check the Research Ethics page on the College of Sciences & Engineering (CoSE) Health and Safety Services website.* [↑](#footnote-ref-1)