Link epigenetics to diseases

Using data analytics, how would you propose to improve our understanding of DNA methylation mechanisms and signature for target identification from the 1,000 Genomes Project cohort?

Summary

We invite scientists from all over the world to submit their proposals that demonstrate the usage of our large population cohort data to detect novel methylation patterns in the context of diseases. Submit your research plans no later than **September 19, 2024, 11:59 pm PST**.

What is the context of the problem that we would like to solve?

Technical advances in genome sequencing, including long read sequencing, became very popular in recent years allowing us to overcome technical challenges in linking disease phenotypes to nucleotide polymorphisms. Our goal is to stimulate novel research on DNA methylation using our data to improve the understanding of disease mechanisms.

For the first time, we are offering a collection of approximately 900 human samples from the 1,000 Genomes Project sequenced using Oxford Nanopore Technology (ONT) free of charge and sharing it with data scientists across the globe to foster open science. Data will be shared based on an open access creative commons license (CC BY-NC 4.0). DNA methylation is an epigenetic mechanism used in cells to control gene expression. It is an epigenetic mark that occurs in cytosines at CpG dinucleotides and affects gene expression under normal physiological conditions and in disease states. Currently available data sets are either of small size or based on short read sequencing methods, limiting our ability to accurately associate methylation states with quantitative trait loci (QTL). Most DNA methylation datasets cannot distinguish between methylation and hydroxy-methylation, two major methylation states. For this, we generated ONT reads from the 1,000 Genomes Project collection making this dataset unique of its kind. Using the appropriate basecallers, DNA methylation and hydroxy-

methylation patterns can be accurately detected, at scale, to address specific questions related to both types of DNA methylations, linking them to quantitative trail loci and pathological mechanisms.

The resulting data package has a size of 550 TB and will be shared in one run, which makes an appropriate data-infrastructure essential. Access to an appropriate data management infrastructure to host a large data set and to process it will be a requisite to participate in this call.

We invite you to send us your proposals using our submission template, where you describe your research purpose focusing on understanding DNA methylation mechanisms and signature for novel target identification. Projects that will start within the next twelve months will be prioritized.

Boehringer Ingelheim will review all proposals and, upon its selection, based on key success criteria outlined as part of this call, provide a limited number of successful applications with access to this exclusive dataset for their independent research. Please note, as part of this call, no additional budgetary funding will be provided. All intellectual property generated from this call will be owned by the respective scientists.

In summary, we look for novel hypotheses that improve our understanding of DNA methylation mechanisms and signature for target identification from the 1,000 Genomes Project cohort.

What potential solutions could be in scope?

The following research proposals using computational approaches are in scope:

- Linking methylation profiles with genetic signatures in the context of human diseases
- Characterizing methylation profiles across ancestries
- Focusing on translational biology for target identification and patients' stratification
- Analyses must start within the next 12 months after the final submission date
- Solutions that agree to use the material according to the creative commons license agreement (CC-BY-NC 4.0).

What potential solutions would be out of scope?

- Requests without research objectives and definite analysis plan within the next 12 months after the final submission date
- Lack of infrastructure that allow for a decentralized storage and archiving of the data package (550 TB)
- Requests with any commercial intent and usage
- Analyses that do not focus on human analyses are out of scope (aiming at using the dataset for secondary or validation analysis)

What benefits do we offer to you in exchange for having submitted a solution?

- Access a large and unique dataset of long read sequences generated using ONT for your research.
- Receive ready-to-use data for the identification of methylation profiles in large cohorts.
- Usage of the data is free of charge. Initial data transfer cost will be paid by Boehringer Ingelheim.
- The material can be freely shared, redistributed, transformed, built upon, and adapted for any purpose, but not for any commercial use according to CC-BY-NC 4.0.
- Own any rights on your results, publish your results independently, and accelerate your research.

What are the key success criteria on which we base our selection for the best answer?

- A well-structured proposal outlining a novel approach that includes usage of our data.
- The proposal should contain a non-confidential description of your research objectives, proposed analysis, anticipated timelines, and planned start of the project.
- Any analyses using the data library must commence within the next 12 months.
- Access to long term storage with capacity for data sharing and backup.
- Preference may be given to application areas aligning with indication priorities of Boehringer Ingelheim outlined in the Supplementary Section, though selection will not be restricted to these areas.
- Proposals that further dissemination of the data publicly will be prioritized.

- Any proposals focusing on target identification and disease understanding would be preferred.

Please note that data are shared under a creative commons license (CC-BY-NC 4.0), i.e., any resulting publications should acknowledge Boehringer Ingelheim as the owner of the data. It would be appreciated to involve Boehringer Ingelheim as co-authors.

Anticipated Project Phases or Project Plan

Phase 1	Review of Proposals will start after the deadline, and we aim to finalize our review by end of October 2024
Phase 2	The data packages will be provided after the decision of our review team

Submitting a collaboration proposal

- Check the 1,000 Genomes Project profile on opnMe or alternatively,
- Click the "Download your submission template" banner to access the collaboration submission template (requires login or registration).
- Follow the instructions to download the template or upload your submission document.
- The upload allows you to attach additional application files if you want to.
- You will be able to access your final submitted collaboration proposal in your personal dashboard and follow its review status.
- Please also visit the FAQ section on opnMe.com to learn more about our Molecules for Collaboration program.

References

1. Noyvert B., Erzurumluoglu A. M., Drichel D., Omland S., Andlauer T. F. M., Mueller S., Sennels L., Becker C., Kantorovich A., Brænne I., Bartholdy B. A., Belbin G. M., Li J. H., Pickrell J. K., de Jong J., Arora J., Kriegl J., Podduturi N., Jensen J. N., Stutzki J., Ding Z. Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. *medRxiv*. **2023**. DOI:10.1101/2023.12.20.23300308.