February 2020

Hello and thank you for taking the time to read the GeCIP newsletter! I hope 2020 is treating you all well so far.
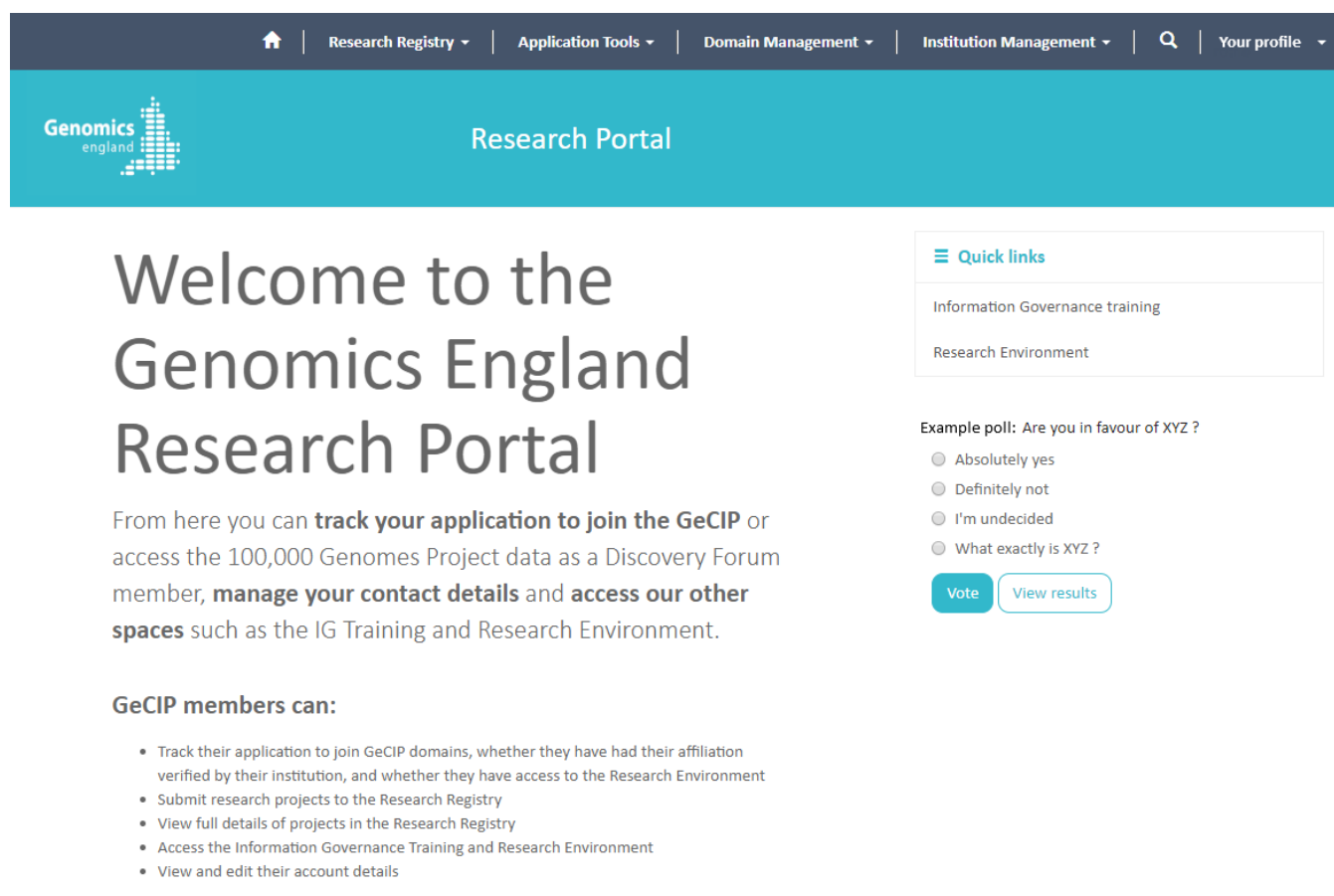
In this quarter's newsletter, you will read about the TRE 2.0 – a new and improved version of our Trusted Research Environment that we will be developing over the coming months. Related to this effort, we will be sending out a brief questionnaire to any of you that are able to access the Research Environment to find out about your experience of it, or - if you have no experience of it, why you don't use it. The survey should take about one minute so please do fill it in if you have the chance.

You will also find information about our new Research Portal and Research Registry, new data, updates and pending changes to the Research Environment, GeCIP publications, how to export data for GeneMatcher, and an exciting genomics meeting in Scotland this summer! As always, if you have questions that this newsletter doesn't answer, please don't hesitate to contact the GeCIP team!

Warmest wishes,

Anna Need

New Research Portal and Research Registry

We are aiming to launch a brand-new online portal and research registry in March/April 2020. The Research Portal will enable you to log on and view your membership of GeCIP domains, track your applications to new domains, update your personal information and contact domain leads. We are aware that for many the GeCIP application process has been somewhat opaque, leaving you unaware of how far along you are in the process, why it is taking so long or if you need to do something. We are pleased to say that this new system will solve these problems, and help researchers gain access to the data more quickly. The portal will also act as a central hub for GeCIP members to access important resources such as links to our Research Environment user guide, your Information Governance training, the Research Environment and *the new Research Registry.*

Within the Research Portal you will find the new Research Registry, which will replace the current Research Registry hosted on Confluence inside the Research Environment. As the new Registry is hosted outside the Research Environment, all GeCIP members – not just those eligible for data access – will be able to browse the projects.

The new Registry will be far easier to browse, with more sophisticated filtering and searching to help you view existing research projects and find potential collaborations. You will be able to draft and save projects before submission, tag your project with genes, phenotypes and keywords, and update or amend the details of your project after submission. There will also be a new section to allow you to update paper publication details. This helps us to showcase your research to both participants and funders. This new system will streamline the project review process and enable researchers to start their work even sooner.

Re-identifying participants and approaching recruiting clinicians

We would like to remind you that under no circumstances are you permitted to re-identify participants in the Research Environment or directly approach recruiting clinicians. Rather, you must use the *Contact Recruiting Clinician* form inside the Research Environment. Even if you personally know the recruiting clinician, it is essential you use this form in order to follow the appropriate governance and maintain the security of this dataset. Full details on how to use this form or submit potential diagnoses can be found here.

Things you may not know about the Research Environment

The Aggregate VCF

We have merged germline genomic (gVCF) data from 59,464 participants from the Main Programme data release 5.1 and have made it available in the Research Environment as a set of multi-sample VCF files for easier handling. The aggregated dataset contains samples from both the rare disease and the cancer programs, but only genomes on build GRCh38 were included. All included samples have passed a set of basic QC metrics and have been annotated for variant quality metrics. Sample QC metrics and other stats are provided in the LabKey table aggregate_gvcf_sample_stats.

The data set is stored within the Research Environment at /gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228/data/, with more information available in the Research Environment User Guide here: https://cnfl.extge.co.uk/pages/viewpage.action?pageId=113194986 and a README document at /gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228/docs/Aggregated_gVCF_dataset_v1_2019_02_28_README.pdf

A second aggregate VCF, consisting of participants from the Main Programme data release 8, is currently being created. We aim to make this available with the upcoming data release 9, followed by a set of VEP-derived annotations for all variants.

Interactive Variant Analysis Update

The Interactive Variant Analysis (IVA) tool allows users to explore individual genomes and aggregate data from the 100,000 genomes project, to identify potentially deleterious variants using filters such as population frequency, consequence type, mode of inheritance and phenotype associations. We are pleased to announce a patch release of IVA, from version 1.0.3 to 1.0.4, which includes:

- New landing page (and additional documentation) to raise user awareness of available functions
- Affected status now shown in the family pedigrees displayed in the interpretation portal
- Adding large parts of GO no longer crashes search (term limit of 100 applied)
- Feature IDs field in 'Genomic' filters no longer freezes for extended periods
- General performance improvements from code optimisation

We are also preparing a major release of IVA, moving to version 2.0.0, by the end of March. This new version will add a range of analytical and other functionality, including:

- Cohort builder to enable the analyses of specific sets of genomes
- Knockout analysis (samples with a KO, and KOs in a sample)
- A range of other new analyses including GWAS
- Support for user-generated custom analyses
- Redesigned interfaces to improve usability

For more information, please consult the roadmaps for IVA and OpenCGA.

RE Feedback Form

We have introduced a new ticket type to our Service Desk specifically for collecting feedback for the Research Environment. If you have any general feedback about the Research Environment, positive or negative, please submit a ticket on the Service Desk Portal by clicking *I'm a GeCIP member*, then *Feedback*. Genomics England will be reviewing the feedback regularly.

Account and data migration

Current issues

We are sorry for those of you who have experienced frustrating instability in the Research Environment. This has come as a result of our data storage migration (see below) that will ultimately make your analyses faster and we expect this issue to be resolved this spring.

HPC Queue

Due to a misconfiguration in the priority scheduling system, jobs submitted to the long and medium queues have been run on a first-come-first-serve rather than a fair share basis in recent weeks. This has now been rectified. Apologies for the inconvenience and delay caused by this and thank you to the users that helped us identify this issue.

New account system

We have changed to a new authenticator provider, so you will have received an email to reset your Research Environment password. If you have not received your password or are having any difficulties logging in since this change, please contact our Service Desk.

Research Environment storage migration

We are commencing a migration of all genomes in the Research Environment to a new storage system. The migration will last approximately two months. Each genome will be temporarily unavailable when it is migrated. This means that, in practice, any HPC jobs trying to access a genome at the time of its migration will fail. In most cases, as each genome will only be unavailable for a couple of minutes, restarting your job will resolve the issue. However, jobs trying to access multiple genomes will be more severely affected. While this will cause some disruption over the next two months, once the migration is completed, the high-performance compute cluster will run significantly faster.

It is therefore recommended that you consider adding extra file-checking steps to your code so that it fails gracefully if a file is unavailable. If your code is looping over many genome files, you might consider modifying it such that it is able to skip any files that appear to be unavailable, while keeping track of them so that they can be tried again later.

Research Environment usage survey

Many of you will have received an email asking you to answer a few short questions based on your activity in the Research Environment. Please fill out this survey, even if you have never used the Research Environment, as all your feedback will be critical in shaping its further development. We are aiming for 100% completion rate, and will be sending reminder emails to those who haven't yet filled it in.

Registering Research Projects

We would like to remind GeCIP members of the importance of registering projects in the Research Registry. **All research carried out within the Research Environment must be aligned with a project in the Research Registry.** The Research Registry is the primary way in which the GeCIP Team and domain leads identify overlaps and prevent the duplication of efforts. Registering your project is also necessary in order to take out any results via the airlock. We would expect each registered project to correspond roughly to a single research paper.

You can find the Research Registry by logging into the Research Environment and clicking the Research Registry icon on the desktop. Here you can view the current projects and register your own. You will need to write a short abstract and lay summary for your project. Please be aware that at the request of the Participant Panel, projects may be rejected if the lay summary is found to be too brief or too technical. If you would like to amend any details or add contributors to an existing project, please email the GeCIP Team at gecip-help@genomicsengland.co.uk. Click here to read the Research Registry user guide.

Recent GeCIP publications

Since our last newsletter in November, we have been informed of 19 new publications based on the 100,000 Genomes Project. These include submissions from Siddharth Banka's group on variants in KMT2D as well as the co-lead of the Ethics and Social Science domain, Anneke Lucassen, exploring the consent of the 100,000 Genomes Project.This takes the total number of 100,000 Genomes Project publications we have been informed about up to 92!You can find all of these publications on the publications page of our website. If you are working on a publication on the data of the 100,000 Genomes Project, please get in touch with the GeCIP team so we can add this to our records. It is essential for the funders, thanks to whom the Research Environment is free to access, that publications and any research outputs from the 100,000 Genomes Project are tracked. We will also publicise it on the Genomics England website and may even tweet about it. These publications are the main way participants see and understand the research being done on their genomes, so we owe it to them to make these publications as visible as possible.

How to correctly acknowledge Genomics England in publications

We have received questions from GeCIP members on how to correctly acknowledge Genomics England and when it is necessary. These acknowledgements are important because funders are looking for them in assessing the impact of the 100,000 Genomes Project and the utility of the Research Environment.When a publication has used 100,000 Genomes Project data or contacted participants through Genomics England, it is essential you follow our full requirements outlined on our publications page. This entails (1) sending your manuscript to the GeCIP Team (gecip-help@genomicsengland.co.uk) two weeks in advance of submission, (2) including the Genomics England Research Consortium as an author and (3) using our official acknowledgement texts for papers, abstracts and posters. Full details are available on our website here. For publications not using Project data, but still related to the 100,000 Genomes Project or Genomics England, while not obligatory we do request that you use the same acknowledgement text.. Ultimately this kind of research would not have been possible without the Project, therefore these publications should be likewise recognised as part of the Project's impact.We will be updating our publications page with more detailed guidance on all research scenarios and their corresponding acknowledgement and citation requirements, so please refer to this when you intend to publish.

GeneMatcher form in Airlock

We know many of you are keen to submit genes to GeneMatcher based on your findings from the 100,000 Genomes Project in the Research Environment. In order for us to maintain a record of what has been submitted, prevent duplication and monitor what leaves the RE, please ensure that any submissions to GeneMatcher align with one of your registered research projects and are taken through the airlock. We have a new airlock form available specifically for this and it should be a very quick and easy process. It will only require review by the Airlock Manager, not the committee. Feedback welcome.

Genomic Medicine: Moving beyond the sequence – June 29th 2020

Genomics England and the British Society for Genetic Medicine, in association with the Scottish Genomes Partnership, are pleased to announce that **registration is now open for the "Genomic Medicine - Moving Beyond the Sequence" meeting**. Register now at: https://gm-beyond-the-sequence. eventbritestudio.com, where further details of the meeting are available.In tandem with the main meeting, we are also hosting two satellite workshops: "Genetics, Genomics and the Law" (June 29th) which addresses topical areas of overlap between genetics and the law including the current ABC case; and "Genomic Medicine - Health Economics" (July 1st), which will address the affordability and benefits of genome-based analysis and testing. Registration for the two satellites is also open via the above web page.Thanks to sponsorship by Genomics England, BSGM, SGP and other sponsors, we have been able to keep registration fees low: £20 for the main meeting and £10 for each of the two satellites. We hope you will be able to join us for this summer visit to Edinburgh!