UKBC

Conference of Bioinformatics and Computational Biology

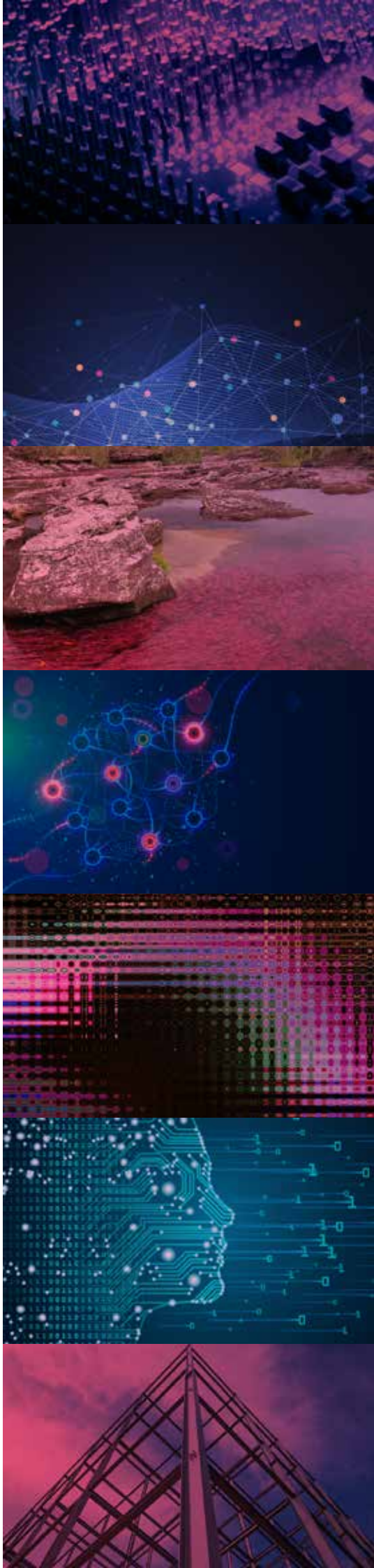
27 - 29 September 2022

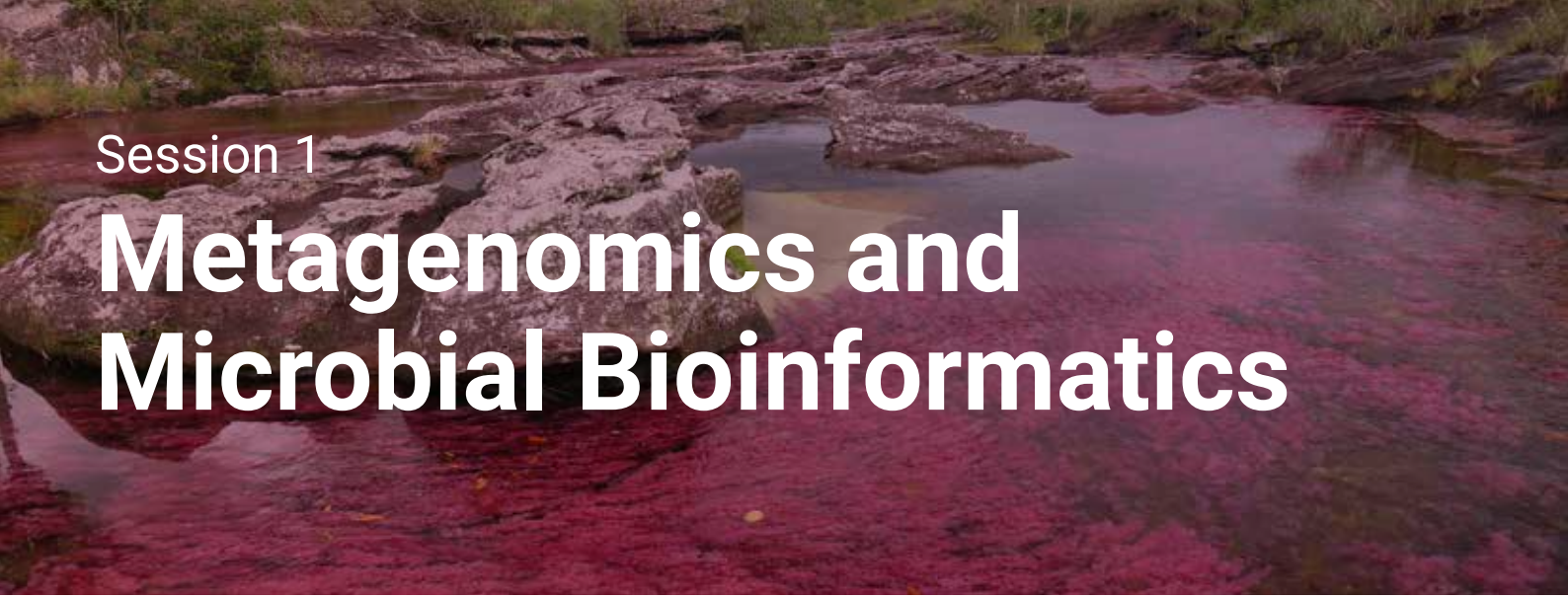
Abstract Booklet

In association with

Earlham Institute

elixir UNITED KINGDOM





Session 1

Metagenomics and Microbial Bioinformatics

Session Facilitator

Rob Finn
European Bioinformatics Institute (EMBL-EBI)

Metagenomics is the study of the sum of genetic material from the microbes living in any particular environment.

Metagenomics datasets represent important spatial and temporal records of microbes along with additional factors, which can range from disease state to abiotic factors, such as pH or salinity. The field of metagenomics has burgeoned due to increased access and advances in modern sequencing technologies coupled with the ever-diminishing costs associated with these approaches.

Consequently, researchers are generating and analysing huge datasets at scale, which demonstrate great potential for a multitude of diverse applications. However, there are major informatics challenges in aggregating, analysing, and comparing data. In this session, we will discuss approaches to address these challenges and identify the major hurdles and gaps in the current repertoire of informatics tools.

Importantly, we will aim to address the principal elements that constitute a dataset, the reuse of corresponding analysis results as well as determine the key data products required by researchers. Furthermore, we will also consider the periodic updates of analysis over time. Finally, we will consider emerging sequence technologies and other approaches (e.g. metabolomics) that impact microbiome analysis. The speakers in this session traverse the spectrum from algorithm developers to data repositories and downstream analysis.

Join the conversation: **#UKCBCB** @EarlhamInst @ElixirNodeUk

UKCBCB
Conference of Bioinformatics and Computational Biology

Session 1 | Talk 1

Luiz Irber
Computational Biologist
10x Genomics

Talk

Content-Based Petabase-Scale Search with Fractional Sketches

Abstract

Public sequencing databases like the SRA contain tens of petabytes of data, but are largely unsearchable due to their large size and continuous growth.

Most of the deposited studies looked into specific questions, but they can still reveal patterns for questions that were not considered initially.

This talk will present methods for working with content-based search of public sequencing databases, and some ideas for related large scale data analysis with the results, with initial focus on biogeography and outbreak tracking.

Session 1 | Talk 2

Robert Griffiths
UK Centre for Ecology and Hydrology, Bangor

Talk

Genes across landscapes: ecological approaches toward synthesis of microbiome diversity & function

Abstract

Advances in sequencing technologies over the past decade have seen an enormous rise in microbial biodiversity (16S) datasets from diverse habitats.

This talk discusses how we can use large scale data to bring knowledge synthesis and further practical applications of “microbiome” data.

Pressing needs include i) synthesis on the ecology of new phylotypes ii) formalising functional linkages, and iii) testing of predictions. These issues will be discussed with respect to recent work on soils, though the concepts should be applicable to other systems.

Specifically I will show the linking of various datastreams to generate niche-models of individual phylotypes, which predict community level change, and inform on how environmental perturbation affects diverse taxa within different environmental contexts.

I further discuss how whole genome metagenomics can be used to complete evidence chains linking environmental drivers to taxon responses and community functional potential.

Finally I propose that new digital solutions are required to synthesise coupled genomic and ecological information, to facilitate a transition from microbiome discovery toward microbiome hypothesis testing.

Session 1 | Talk 3

Lorna Richardson
European Bioinformatics Institute
(EMBL-EBI)

Talk

MGnify data products for reference-based analysis and metagenomic mining

Abstract

MGnify provides consistent pipeline-based analyses of sequence-based microbiome data, including amplicon, metagenomic, metatranscriptomic and assembly data. We provide metagenomic assembly as a service, depositing all assembled datasets into public repositories.

Two main data products have developed from this large-scale effort of assembling data, namely a collection of biome-specific genome catalogues, and a database of 2.4bn non-redundant protein sequences.

These data products provide a framework for researchers to determine novelty within their own microbiome datasets, as well as to allow mining of the data for enzymatic functions of interest.

Session 2

Bioimaging and Artificial Intelligence

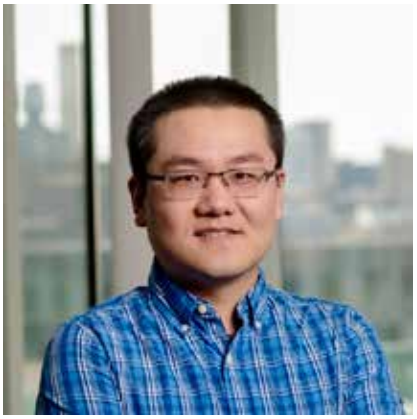
Session Facilitator

Jean-Marie Burel
Senior Software Architect
University of Dundee

Bioimaging measures the structure, constitution and dynamics of molecules, cells, tissues and organisms in biological systems. Data are often GBytes or Tbytes in size, cover space, time and other dimensions, include complex metadata describing experimental setups, acquisition parameters and analytic outputs and are stored in 100s of different proprietary file formats.

The scale and complexity of these data are a consistent, unsolved block to researchers making their bioimaging data public and FAIR, despite the fact that open data sharing and publication are mandated by funding agencies.

To address some of those points, we will discuss the on-going work on a new community-agreed cloud-friendly FAIR file format, OME-NGFF. We will discuss existing public bioimaging repositories (BIA, IDR) and how to make FAIR bioimage data possible. Finally, we will introduce new software tool, using public imaging resources, that integrates heterogeneous datasets enabling hypothesis generation.



Session 2 | Talk 1

Yang Zhang
Project Scientist & Project Manager,
Carnegie Mellon University

Talk

Nucleome Browser: An integrative and multimodal data navigation platform for 4D Nucleome

Abstract

We introduce Nucleome Browser (http://www.nucleome.org), an interactive, multimodal data visualization and exploration platform for 4D Nucleome research.

Our tool effectively integrates heterogeneous datasets (e.g., genomics, imaging, 3D genome structure models, and single-cell data) and external data portals by a new adaptive communication mechanism.

Nucleome Browser provides a scalable solution for integrating massive amounts of 4D Nucleome data to navigate multiscale nuclear structure and function in a wide range of biological contexts, enabling hypothesis generation and data sharing with the broad community.

Session 2 | Talk 2

Matthew Hartley
BioImage Archive Team Leader,
EMBL-EBI

Talk

Open bioimaging data at scale:
publication, analysis and reuse

Abstract

Imaging data plays a huge role in driving biological research.

Scientific results are often generated by complex data analysis pipelines that operate on images to produce insights, with AI approaches playing an increasing role in this process. However, without providing access to the raw data that feeds these data pipelines, we risk a future reproducibility crisis.

In this talk, I will discuss how EMBL-EBI's imaging data resources, in particular the BioImage Archive support reproducibility through allowing broad access deposition of any imaging dataset.

I'll also explain how providing FAIR imaging data at scale can drive data reuse and methods development, particularly in AI.

Session 2 | Talk 3

Josh Moore
Senior Software Architect,
University of Dundee

Talk

OME-NGFF (next-generation file format):
Zarr as a cloud-native solution for FAIRer bioimaging data

Abstract

To date, the rapid innovation in biological imaging and diversity of applications have prevent the establishment of a community-agreed standardized data format.

The Open Microscopy Environment (OME) -- an open-source software project that develops tools that enable access, analysis, visualization, sharing and publication of biological image data – is leading a community effort to design a new cloud-friendly FAIR file format — OME-NGFF which, together with existing open formats like OME-TIFF and HDF5, should satisfy the majority of bioimaging use cases.

Such next-generation file formats (NGFFs) like Zarr are characterized by the use of individual compressed chunks which facilitate the parallelization of creation and access.

As a part of the development of OME-NGFF, we are converting sample datasets from Image Data Resource (IDR) and making them publicly available on S3 storage as well as providing tools to convert existing imaging data to OME-NGFF and creating libraries for the visualization of chunked formats in diverse ecosystems like Fiji, napari, and Viv.

Session 3

Sex and Gender Bias in Computational Disciplines

Session Facilitator

Franca Fraternali
Randall Centre for Cell and Molecular Biophysics, King's College London

There are multiple sources of undesirable biases that impact computational disciplines, from underrepresented groups in the research body (females and minorities) to unbalanced data sources that are used for machine learning and decision processes in Life Sciences.

The effect is a skewed distribution of stakeholders in the development of research in computational disciplines and a resulting need for additional data curation to arrive at balanced and truly representative data.

We want to generate a constructive debate in the Computational, Bioinformatic and Biomedical communities to analyse social, ethical and technological challenges resulting from these biases.

Session 3 | Talk 1

Alba Jené Sanz
Bioinformatics Unit Coordinator,
Barcelona Supercomputing Centre

Talk

The Bioinfo4Women programme: promoting gender equity and diversity for science at the BSC

Abstract

The Bioinfo4Women programme (B4W) is an initiative that started in 2018 to promote the research done by women in computational biology, and it supports researchers by promoting the exchange of knowledge and experience of outstanding women researchers through activities such as seminars, conferences, training and mentorships.

B4W has particular focus on the areas of personalised medicine, bioinformatics and HPC, and ultimately aims at building a more collaborative, supportive, and equal scientific community.

Here, we will showcase the activities of B4W in three fronts: 1) summary of the research line on sex and gender biases in AI; 2) training to raise awareness on the matter to early stage researchers; and 3) pilot international mentoring programme to provide role models to early stage scientists with a gender and diversity perspective.

Session 3 | Talk 2

Elena Bernabeu
PostDoctoral Fellow, Institute of Genetics and Cancer, University of Edinburgh

Talk

Unravelling the effect of sex on human genetic architecture

Abstract

Males and females present differences in complex traits and in the risk of a wide array of diseases. Genotype by sex (GxS) interactions are thought to account for some of these differences.

However, the extent and basis of GxS are poorly understood. In the present study, we provide insights into both the scope and the mechanism of GxS across the genome of about 450,000 individuals of European ancestry and 530 complex traits in the UK Biobank.

We found small yet widespread differences in genetic architecture across traits. We also found that, in some cases, sex-agnostic analyses may be missing trait-associated loci and looked into possible improvements in the prediction of high-level phenotypes.

Finally, we studied the potential functional role of the differences observed through sex-biased gene expression and gene-level analyses. Our results suggest the need to consider sex-aware analyses for future studies to shed light onto possible sex-specific molecular mechanisms.

Session 3 | Talk 3

Segun Fatumo
Associate Professor, MRC/UVRI Uganda and London School of Hygiene and Tropical Medicine

Talk

Ethnic bias in genomics studies

Abstract

Since the sequence of the first human genome about twenty years ago, there have been advances in genome technologies which have resulted in whole-genome sequencing and microarray-based genotyping of millions of human genomes. However, genetic and genomic studies are predominantly based on populations of European ancestry.

For example, as of June 2021, the vast majority of genomics studies including genome-wide association studies (GWAS) have been conducted in individuals of European descent (86.3%), followed by East Asian (5.9%), African (1.1%) populations. While the proportion of samples from individuals of European ancestry has increased from 81% in 2016 to 86% in 2021, the proportion of samples from the underrepresented populations have either stagnated or decreased.

As a result, the potential benefits of genomic research including better understanding of disease etiology, early detection and diagnosis, rational drug design and improved clinical care may elude the many underrepresented populations.

In my presentation, I will demonstrate the value of genomic diversity and how we are using limited Africa human genome resource for gene discovery and genetic risk prediction.

Session 4

Federated Analytics / Learning

Session Facilitator

Phil Quinlan
Honorary Professor and Director of Health Informatics, University of Nottingham

Federated Analytics and learning have been around for a while, and whilst not seeking to open the debate about centralised vs federated models, there are scenarios where it is desirable not to move data to a single location.

The session will hear about work in the UK to bring in a national platform to allow discovery questions to be asked across Trusted Research Environments. This will be followed by a presentation on some of the issues that can exist once we move from simple counts to more advanced analytics. Data usually resides in a Trusted Research Environment to prevent disclosure of data but can the algorithms themselves disclose information about the participants.

Finally, we will hear about data shield that has developed R libraries that are compatible with some disclosure control concerns and has been in practice in both organisational and life science datasets.

The breakout sessions will cover:
(1) the understanding of federated systems and which have been used successfully
(2) the understanding of disclosure control and where techniques have been deployed
(3) standards and interoperability between systems.

Join the conversation: **#UKCBCB** @EarlhamInst @ElixirNodeUk

UKCBC
Conference of Bioinformatics and Computational Biology

Session 4 | Talk 1

Thomas Giles
Team Leader, The Digital Research Service,
The University of Nottingham

Talk

Tools to enable rapid discoverability of secure healthcare datasets (CaRROT and HUTCH)

Abstract

One of the major blockers in making healthcare data FAIR is the resource required to convert data to a common data format, in this talk we present CaRROT; a low risk, GDPR-compliant tool that allows remote data mapping by only operating on metadata that is outside of the scope of data governance regulations.

This allows teams with specialist expertise in data mapping to assist data partners in undertaking an ETL (Extract, Translate and Load) process to a common data format without ever having access to the underlying data.

When making data discoverable via federated approaches another potential blocker is the interoperability of the local component that sits within a data partners infrastructure with the various data and metadata portals (both existing and in developemnent).

In this talk we will also present HUTCH, an interoperable and customisable infrastructure layer tool to make secure healthcare data that, due to governance constraints, cannot be shared without approval, discoverable.

Session 4 | Talk 2

Emily Jefferson
Health Informatics Centre (HIC) Director,
University of Dundee

Talk

Recommendations for disclosure control of trained Machine Learning (ML) models from Trusted Research Environments (TREs)

Abstract

Trusted Research Environments (TREs) are widely used to support statistical analysis of sensitive data across a range of sectors (e.g., health, police, tax and education) as they enable secure and transparent research whilst protecting data confidentiality.

There is an increasing desire from academia and industry to train AI models in TREs. The field of AI is developing quickly with applications including spotting human errors, streamlining processes, task automation and decision support.

These complex AI models require more information to describe and reproduce, increasing the possibility that sensitive personal data can be inferred from such descriptions. TREs do not have mature processes and controls against these risks. This is a complex topic, and it is unreasonable to expect all TREs to be aware of all risks or that TRE researchers have addressed these risks in AI-specific training.

This talk describes the work of the GRAIMATTER project which has developed a setof usable recommendations for TREs, researchers and data controllers to guard against the additional risks when disclosing trained AI models from TREs.

Session 4 | Talk 3

Hugh Garner
Full Stack Developer
Newcastle University

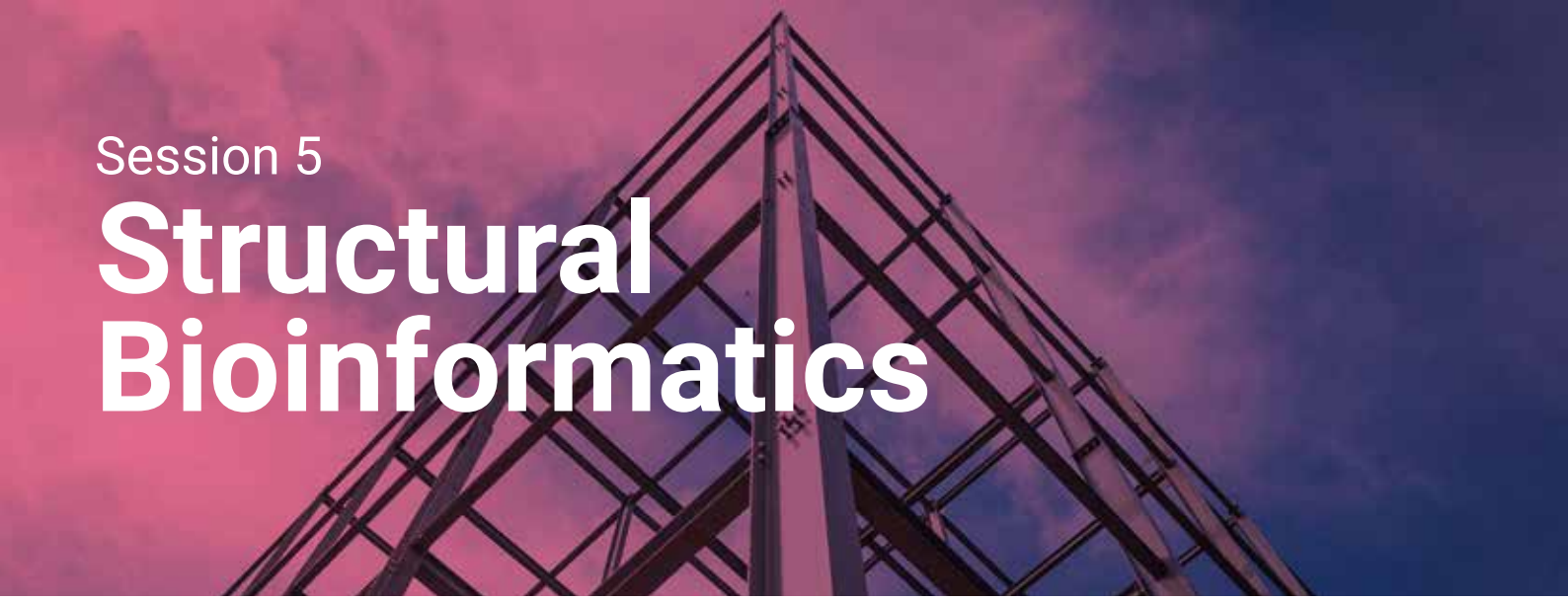
Talk

Privacy-preserving federated analysis with DataSHIELD: methods and experiences

Abstract

DataSHIELD is a widely used set of infrastructure and R packages that enable the federated non-disclosive analysis of sensitive biomedical data.

This talk covers the reasons for developing DataSHIELD, technical and statistical methodology including disclosure controls. Some key user case studies from large consortia are highlighted, plus lessons learned from deployment and development of the DataSHIELD ecosystem over the last 8 years.



Session 5

Structural Bioinformatics

Session Facilitator

Mark Waas
School of Biosciences,
University of Kent

We're in an exciting era for structural bioinformatics, where AlphaFold2 has enabled the structure of the many millions of proteins to be modelled accurately.

This session will consider the diverse areas of structural bioinformatics, including protein evolution, analysis of genetic variants, and also the protein structures encoded by alternative splice isoforms. We will consider the opportunities that AlphaFold2 offers for research in each of these areas.

Session 5 | Talk 1

Joe Marsh
Group Leader, MRC Human Genetics Unit,
University of Edinburgh

Talk

Interpreting variant effects through the lens of protein structure

Abstract

Most known pathogenic mutations occur in protein-coding regions of DNA and change the way proteins are made. Taking protein structure into account has therefore provided great insight into the molecular mechanisms underlying human genetic disease.

While there has been much focus on how mutations can disrupt protein structure and thus cause a loss of function (LOF), alternative mechanisms, specifically dominant-negative (DN) and gain-of-function (GOF) effects, are less understood.

In recent work, we investigate the protein-level effects of pathogenic missense mutations associated with different molecular mechanisms. We observed striking differences between recessive vs dominant, and LOF vs non-LOF mutations, with dominant, non-LOF disease mutations having much milder effects on protein structure, and DN mutations being highly enriched at protein interfaces.

We also observed that, while nearly all computational variant effect predictor underperform on non-LOF mutations, consideration of their tendency to cluster in three-dimensional space may improve their identification.

Session 5 | Talk 2

Yana Bromberg
Professor, Rutgers University School
of Environmental and Biological Science

Talk

Protein fossils: metal binding then and now

Abstract

Biological redox reactions drive planetary biogeochemical cycles.

Using a novel, structure-guided sequence analysis of proteins, we explored the patterns of evolution of enzymes responsible for these reactions.

Our analysis reveals that the folds that bind transition metal–containing ligands have similar structural geometry and amino acid sequences across the full diversity of proteins.

Similarity across folds reflects the availability of key transition metals over geological time and strongly suggests that transition metal–ligand binding had a small number of common peptide origins.

We observe that structures central to our similarity network come primarily from oxidoreductases, suggesting that ancestral peptides may have also facilitated electron transfer reactions.

Last, our results reveal that the earliest biologically functional peptides were likely available before the assembly of fully functional protein domains over 3.8 billion years ago.

Session 5 | Talk 3

Michael Tress
Spanish National Cancer Research Centre (CNIO)

Talk

AlphaFold, Model Structures and Genome Annotation

Abstract

The advent of AlphaFold has revolutionized the world of protein structure prediction.

Of particular interest to the genome annotation community is the fact that AlphaFold and related models can make predictions for regions of amino acid sequence that were previously untreatable by prediction methods.

In theory, these new prediction methods could be used to validate protein coding regions, predict changes in structure due to alternative splicing, and even to search for novel folds.

Here, I show what AlphaFold and related methods might achieve, what the limits are, and how to avoid pitfalls, using concrete examples from research into the human genome.

Session 6

Open Science

Session Facilitator

Yo Yehudi
Executive Director,
Open Life Science

As researchers, we may be asked to share our data, code, and workflows, perhaps to meet the needs of a funder or institutional policy. On the other hand, we might be interested in sharing our work, but perhaps there is not a culture of sharing, or privacy concerns prevent transparent data sharing.

Making sure that we share our work effectively, in ways that facilitate re-use, isn't always straightforward.

This session will feature presentations from three open science practitioners, who have experience not only sharing their computational and biological work, but also building communities around their projects through all stages of the research lifecycle.

Our three tracks will offer the chance to talk about practical, real-life concerns around specific biological and open community related questions, including bringing in new community members and users of your data and tools.

Session 6 | Talk 1

Batool Almarzouq
Postdoctoral Researcher, King Abdullah International Medical Research Centre / University of Liverpool

Talk

Leveraging open science in machine learning and bioinformatics

Abstract

Biology has become a rich, data-intensive science, dependent on complex, computational, and statistical methods, where machine learning and deep learning algorithms can be leveraged to provide novel insights for complex biological questions.

Open Sciences has been instrumental in making these methods accessible to researchers while ensuring scientific results remain reproducible.

This talk will discuss how open science practices can boost bioinformatics research and open new avenues for promoting scientific discovery by extending the principle of openness to the whole research cycle. We will explore examples of how Open Science is applied in the field of machine learning and computational biology.

We will look into both the challenges and advantages of applying Open Science practices in the ever-evolving field of machine learning. We aim to prompt attendees to reflect on specific concerns and practices that impact data science and bioinformatics research. We will be using a shared document for collaborative note taking to capture the diverse experience of the participants and derive actions to help more communities adopt open science practices.

Session 6 | Talk 2

Piraveen (Piv) Gopalasingam
Scientific Training Officer,
European Bioinformatics Institute
(EMBL-EBI)

Talk

Open science training and equitable exchanges in global collaborations

Abstract

There is a critical need to address global challenges such as crop security, biodiversity protection, and communicable disease surveillance.

Often these interdisciplinary topics demand global collaborations, requiring cooperation between groups in the Global North and South or High:Low-Middle Income countries (LMICs). Differences in existing resources and capacity permit unethical practices such as helicopter research and perpetuate inequity.

However, training and practising equitable collaboration behaviours can form part of capacity development to transfer skills and strengthen research. There is a worldwide need for scalable, open bioinformatics training and efforts have been made by communities to address this gap, such as CABANA in Latin America, H3ABioNet in Africa and APBioNet in the Asia-Pacific region.

Together, these groups created guidelines for organising and delivering training in LMICs, and through the CABANA project equitable collaboration was promoted, inclusive selection of trainees and secondees were co-developed and later, workshops ran in local languages.

Session 6 | Talk 3

John Ogunsola
School of Biodiversity, One Health
& Veterinary Medicine,
University of Glasgow, UK

Talk

Global Distribution of APOL1 Genetic variants: a case study into equitable measurement of global genomics

Abstract

The human phenotype is a consequence of underlying genetic variation and environmental influences. With newer technologies, such as next-generation genomic sequencing, scientists are beginning to unravel how genetic variation results in susceptibility to several communicable and non-communicable diseases globally.

Lately, there have been calls to make these genomic datasets open and publicly available. Despite accounting for 20% of the world's population, open genomic data from Africa does not match equitably with those from other parts of the world.

Using genetic variants in a protein that have been shown to increase the risk of non-diabetic chronic kidney disease as a case study, this talk will demonstrate the inequality in measurement of global genomics, raise salient points as to direct and indirect contributory factors, and provide a platform for discussion on how this gap might be bridged.

Session 7

Spatial Transcriptomics

Session Facilitator

Sarah Teichmann
Head of Cellular Genetics and Senior Group Leader, Wellcome Sanger Institute

As researchers, we may be asked to share our data, code, and workflows, perhaps to meet the needs of a funder or institutional policy. On the other hand, we might be interested in sharing our work, but perhaps there is not a culture of sharing, or privacy concerns prevent transparent data sharing.

Making sure that we share our work effectively, in ways that facilitate re-use, isn't always straightforward.

This session will feature presentations from three open science practitioners, who have experience not only sharing their computational and biological work, but also building communities around their projects through all stages of the research lifecycle.

Our three tracks will offer the chance to talk about practical, real-life concerns around specific biological and open community related questions, including bringing in new community members and users of your data and tools.

Session 7 | Talk 1

Catalina Vallejos
Group Leader, MRC Human Genetics Unit at University of Edinburgh and The Alan Turing Institute

Talk

Scalable Bayesian methods to robustly quantify cell-to-cell molecular variability

Abstract

Cell-to-cell variability in seemingly homogeneous cell populations plays a crucial role in tissue function and development. Single-cell sequencing technologies can characterise this variability in an unbiased manner but its output is prone to high levels of technical noise and sparsity.

This has motivated the development of new bespoke statistical methods to robustly unlock rich knowledge from these complex data. As the field matures, large scale studies using these technologies have become commonplace.

This has introduced additional computational challenges, particularly for high-dimensional models designed under a Bayesian paradigm. I will describe different algorithms that may be used to alleviate this problem.

Using motivating examples (single cell measurements of gene expression and DNA methylation), I will also illustrate how significant improvements in speed may be achieved without substantially sacrificing estimation performance.

Session 7 | Talk 2

Mo Lotfollahi
Helmholtz Munich /
Welcome Sanger institute

Talk

AI to understand health and disease using single-cell atlases

Abstract

The increasing availability of multimodal single-cell technologies provides exciting opportunities to learn a holistic view of cellular behaviours.

Yet, learning and transferring knowledge from multimodal reference atlases remains a fundamental challenge.

I will talk about this challenge and the solutions we proposed using machine learning to construct and facilitate the usage of such an atlas across different data modalities in both health and disease.

Session 7 | Talk 3

Emma Dann
PhD Student in Computational Biology, Cellular Genetics Programme, Wellcome Sanger Institute

Talk

Mapping the developing human immune system across organs

Abstract

Recent advances in single cell genomics technologies have facilitated studies on the developing immune system at unprecedented scale and resolution. However, these studies have focused on one or a few organs and were thus limited in understanding the developing immune system as a distributed network across tissues.

We integrated single-cell RNA sequencing and spatial transcriptomics profiles of prenatal haematopoietic organs, lymphoid organs and peripheral organs to reconstruct the developing human immune system and tissue niches of immune cells.

Using our multiorgan scRNA-seq dataset as a reference, we mapped cell types in tissue to identify niches of immune cells in early hematopoietic tissue and lymphoid organs critical for B and T cell development.

Our analysis uncovered system-wide blood and immune cell development beyond the conventional primary haematopoietic organs, and localization and interactions of immune cell progenitors across tissues.

We provide preprocessed data and pre-trained models for cell embedding and automatic cell type annotation, to streamline future expansion and analysis of human developmental atlases to facilitate cell engineering, regenerative medicine and disease understanding.

Keynote Lecture

Serena Scollen

Head of Human Genomics and Translation Data at ELIXIR

Talk

Towards cross-border access to human genomes at scale for research and healthcare

Abstract