Grand Challenges

2026 Google DeepMind Fellowship in Genomics and AI

wellcome sanger institute

Google DeepMind

Grand Challenges:

2026 Google DeepMind Fellowship in Genomics and AI

Presented below are 6 ‘grand challenges’ representing areas in genomics where AI/ML has great potential to deliver societal impact. The GDM Fellow will tackle 1* of these grand challenges, using their unique AI/ML skills to further the challenge, under the supervision of a Wellcome Sanger Institute faculty member, and with co-supervision from an AI/ML expert at a UK university.**

For the GDM Fellowship purpose, package and eligibility criteria visit the GDM Fellowship webpage

Grand challenge titles	
Grand Challenge 1	Spatial Modelling of Tumour–Immune Co-evolution Driven by Y Chromosome Loss
Grand Challenge 2	Predictive and Mechanistic AI for Longitudinal Multi-Omics in Complex Disease
Grand Challenge 3	Solving the ‘second secret of life’ using massive data generation and machine learning
Grand Challenge 4	Harnessing multimodal generative AI to study the interplay of genetics with spatial transcriptomics, proteomics and morphology in human tissues
Grand Challenge 5	A Foundation sequence model of the Cis-Regulatory Code in developing human tissues
Grand Challenge 6	AI for solving the gene regulatory code across life

**Fellowship candidates may express interest in up to 3 ‘Grand Challenges’ in their application, but will only work one if their application is successful*

***The AI/ML co-supervisor will be determined once the Fellow, and which ‘Grand Challenge’ they will work on, has been finalised*

Grand challenge overviews

Grand Challenge 1: Spatial Modelling of Tumour–Immune Co-evolution Driven by Y Chromosome Loss

Science Programme: Somatic Genomics

Lead Supervisor(s): David Adams - Interim Head of Somatic Genomics and Senior Group Leader

As men age, many of their cells lose their Y chromosome, one of two sex chromosomes determining biological sex. This Loss of Y chromosome (LOY) affects up to 40% of elderly men and strongly links to cancer development and death. Scientists once considered LOY a harmless correlate of aging, but recent discoveries reveal that it plays a role in both cancer and immune cell dysfunction and is associated with poor outcomes. Your immune system should destroy cancer cells, but when immune cells lose their Y chromosome, they may become exhausted and dysfunctional. Further, Y-deficient tumours seem to attract more Y-deficient immune cells, building a protective shield ("contagion hypothesis") from immune-mandated killing.

Through the use of cutting-edge spatial technologies that have profiled tissues at single cell resolution, we can identify each cell's location and gene activity, like a detailed map of each tumour. Artificial Intelligence can analyse 100s of tissue samples including from bladder, lung, kidney, pancreatic, and skin cancers to identify "danger zones" where Y-deficient immune cells cluster around tumours. This ‘Grand Challenge’ seeks to use AI to run virtual experiments, computationally removing or restoring Y chromosomes, to predict whether we can reverse immune exhaustion.

These analyses will transform cancer treatment in aging men by revealing:

1. Why immune systems fail
2. Paths to restoring immune competence

Beyond cancer, understanding Y chromosome loss could explain sex differences in infections and autoimmune diseases. The computational maps will be made freely available worldwide, accelerating discovery.

Grand Challenge 2: Predictive and Mechanistic AI for Longitudinal Multi-Omics in Complex Disease

Science Programme: Human Genetics

Lead Supervisor(s): Carl Anderson - Head of Human Genetics Programme and Senior Group Leader

Modern genomics and healthcare are generating an unprecedented depth of data: longitudinal multi-omics spanning genomes to single cells, combined with rich clinical records capturing how disease unfolds over time. Yet despite this, our ability to predict who will develop disease, which patients progress to severe phenotypes, and how patients will respond to treatment remains limited. Current approaches tend to analyse these data in isolation — snapshots of single modalities or static clinical records — missing the complex, time-dependent biology that drives real-world disease and falling short of the predictive tools needed for precision medicine.

This project addresses that gap by developing next-generation AI models that can integrate high-dimensional, irregular, and multi-modal longitudinal data into predictive, mechanistically grounded insights that enable prediction of disease susceptibility, trajectory and treatment response in real patients. This is a fundamental technical and scientific challenge. Longitudinal multi-omics and clinical data are noisy, incomplete, highly confounded and unevenly sampled over time — conditions under which most current AI methods struggle. Yet solving this would be transformative: enabling earlier identification of high-risk individuals, more precise patient stratification, and truly personalised treatment strategies that anticipate failure before it occurs.

Sanger is one of very few places globally with the data depth and governance to do this properly; this AI fellow can define what “good” looks like for the field. Sanger is uniquely positioned to lead this effort, with access to large-scale multi-omics datasets, deeply characterised longitudinal cohorts, and the infrastructure to securely analyse sensitive genomic and health data at scale. This project would sit at the heart of the HumGen vision, turning rich descriptive datasets into predictive and actionable tools for precision medicine and creating a reusable framework for integrating longitudinal multi-omics with clinical data at scale, while establishing a flagship example of AI beyond imaging — defining best practice for integrating multi-omics and clinical data globally.

The work will also prioritise robustness and transferability across populations, including settings with different data structures and healthcare systems, ensuring relevance to global and LMIC contexts. This will help ensure that models are not only high-performing in one setting, but robust, equitable and generalisable across real-world healthcare environments.

The Fellow would:

- Develop AI methods that learn predictive, mechanistically-anchored models of disease susceptibility, progression and treatment response by integrating longitudinal multi-omics (genome, single-cell and bulk transcriptomics, proteomics, metabolomics) with rich clinical trajectories (EHR, imaging, prescribing, procedures) across complex diseases
- Build and benchmark AI architectures for irregularly sampled, multi-modal, longitudinal data (e.g. sequence models, neural ODEs, graph and foundation models)
- Apply these to one or more Sanger-linked longitudinal cohorts (e.g. in IBD and related immune-mediated diseases) to:
  - predict who develops disease,
  - identify who progresses to severe phenotypes, and
  - forecast drug response, loss of response and adverse events
- Focus on models that are interpretable, highlight causal pathways and cell types, and are portable across populations, including LMIC settings

Grand Challenge 3: Solving the ‘second secret of life’ using massive data generation and machine learning

Science Programme: Generative Genomics

Lead Supervisor(s): Ben Lehner - Head of Generative Genomics and Senior Group Leader

Long-range communication in proteins is termed allostery. Allostery is central to biological regulation, allowing proteins to function as molecular switches (or microprocessors). Allostery also underlies the pathogenicity of many disease-causing mutations and, conversely, the efficacy of many licensed drugs.

Despite being proclaimed by Jacques Monod as ‘the second secret of life’, allostery remains poorly understood and very difficult to predict. A key reason for this is the lack of quantitative, large-scale, and well-calibrated data for model training.

To address this data gap we have developed massively parallel experimental methods to produce the first complete maps of allosteric communication in proteins. We are now applying these methods at scale to structurally and functionally diverse proteins to generate many complete experimental allosteric maps and millions of quantitative biophysical measurements for model training. For protein interactions, the resulting datasets quantify changes in the Gibbs free energy of folding and binding for every mutation in every position in each protein, providing orders of magnitude more data than currently available in the scientific literature.

The goal is to use these uniquely sized quantitative data sets to train the next generation of AI models to predict and generate quantitative protein properties from sequence, including protein stability, binding affinity, binding specificity and selectivity, aggregation and allostery from sequence and structure. These models will enable quantitative protein design, optimisation and engineering, mechanistic interpretation of clinical genetic variants, and the identification of new allosteric sites to target therapeutically, including in proteins currently considered ‘undruggable’.

Further relevant reading:

1. Weng C, Faure AJ, Escobedo A, Lehner B. The energetic and allosteric landscape for KRAS inhibition. Nature. 2024 Feb;626(7999):643-652.
2. Beltran A, Naqvi MM, Faure AJ, Lehner B. The allosteric landscape of the Src kinase. Science Advances. 2026 Feb 13;12(7):eaea2726. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022 Apr;604(7904):175-183.
3. Beltran A, Jiang X, Shen Y, Lehner B. Site-saturation mutagenesis of 500 human protein domains. Nature. 2025 Jan;637(8047):885-894.
4. Escobedo A, Voigt G, Faure AJ, Lehner B. Genetics, energetics, and allostery in proteins with randomized cores and surfaces. Science. 2025 Jul 24;389(6758):eadq3948
5. Faure AJ, Martí-Aranda A, Hidalgo-Carcedo C, Beltran A, Schmiedel JM, Lehner B. The genetic architecture of protein stability. Nature. 2024 Oct;634(8035):995-1003.
6. Hidalgo-Carcedo C, Faure AJ, Martí-Aranda A, Zarin T, Lehner B. Allosteric and energetic remodeling of a PDZ domain by protein domain extensions. Nature Communications. 2026 Feb 19.
7. Arutyunyan A, Seuma M, Faure AJ, Bolognesi B, Lehner B. Massively parallel genetic perturbation suggests the energetic structure of an amyloid-β transition state. Science Advances. 2025 Jun 13;11(24):eadv1422.
8. Thompson M, Martín M, Olmo TS, Rajesh C, Koo PK, Bolognesi B, Lehner B. Massive experimental quantification allows interpretable deep learning of protein aggregation. Science Advances. 2025 May 2;11(18):eadt5111.
9. Mighell TL, Lehner B. A small molecule stabilizer rescues the surface expression of nearly all missense variants in a GPCR. Nature Structural Molecular Biology. 2025 Sep 22. doi: 10.1038/s41594-025-01659-6.
10. Taraneh Zarin, Cristina Hidalgo-Carcedo, Ben Lehner. A complete map of specificity encoding enables reprogramming of a dynamic protein interaction. bioRxiv 2024.04.25.591103.
11. Aina Martí-Aranda, Ben Lehner. The evolution of allostery in a protein family. bioRxiv 2025.06.20.660748
12. Maximilian R. Stammnitz, Ben Lehner. The genetic architecture of an allosteric hormone receptor. bioRxiv 2025.05.30.656975.
13. Taylor L. Mighell, Ben Lehner. GPCR-MAPS: high-resolution functional and allosteric mapping of G protein-coupled receptor activation and bias. bioRxiv 2025.05.30.656974.
14. Xianghua Li, Ben Lehner. TF-MAPS: fast high-resolution functional and allosteric mapping of DNA-binding proteins. bioRxiv 2025.10.20.683418.
15. Xiaotian Liao, Ben Lehner. Allostery is a widespread cause of loss-of-function variant pathogenicity. bioRxiv 2025.06.20.660737.

Grand Challenge 4: Harnessing multimodal generative AI to study the interplay of genetics with spatial transcriptomics, proteomics and morphology in human tissues

Science Programme: Cellular Genomics

Lead Supervisor(s): Mo Lotfollahi - Group Leader; Muzlifah Haniffa - Head of Cellular Genomics, Senior Group Leader and Institute Deputy Director

The aim of this challenge is to develop a cross-modal AI model that connects data on genetic variation with cell states and spatial tissue structure. This model would have broad applications across areas like cancer, infectious diseases and chronic conditions, reshaping how we interpret human biology by bridging genetic variation with molecular, cellular and tissue level-information. The Fellow will unite imaging scientists, spatial biologists, genomic researchers, and AI experts at the Sanger Institute, creating a uniquely collaborative and interdisciplinary environment.

One of the biggest challenges in understanding human biology is that different types of data from tissues, such as histopathology, spatial gene or protein expression, and genetic information, are rarely available from the same sample. Each of these data types capture a different layer of tissue biology, but how they relate to one another, especially in the context of disease, remains unclear. We need models that can connect genetic differences between individuals with changes in tissue structure and the expression of genes and proteins that control how cells behave.

Multimodal AI models can help bridge the gap between different types of tissue data by learning how these layers of information interact. These models can then be used to predict one type of data from another, for example, inferring genetic mutations from spatial transcriptomics, histology images, or protein expression, and vice versa. This marks a major shift in how we model and understand tissue biology. It allows us to “fill in” missing information at scale, such as predicting genetic variation from routine clinical histology samples, and linking these predictions to patient outcomes or treatment responses. This approach could greatly reduce experimental costs and make advanced biological analysis more accessible in low-resource settings. So far, most work in this area has focused on linking tissue morphology with spatial gene expression, but the true potential lies in building models that reflect the central dogma of molecular biology, connecting genetic variation with spatial patterns of RNA and protein expression in tissues.

This ‘Grand Challenge’ will benefit from Sanger’s strengths in large-scale, high-quality cellular and tissue genomics, including spatial transcriptomics, proteomics and sequencing to understand genetic variation. Sanger is uniquely positioned to support this effort through:

- Unique access to deeply phenotyped multimodal datasets, such as skin (Haniffa laboratory), brain (Bayraktar laboratory) and kidney (Mitchell laboratory) and strong links with the Human Tissue Spatial Genomics data generation (over 1B cells by 2026 with spatial readouts from 12 PIs at Sanger). This high-quality data is paired and already available for multimodal model training.
- World-class experience and infrastructure, with access to state-of-the-art sequencing technologies to study genetic variation in health and disease, providing a powerful resource for integration into a multimodal model.
- Sanger’s commitment to building frontier AI capabilities that advance genomics. The Lotfollahi laboratory has spearheaded the development of multimodal AI models and this expertise forms a key stepping stone for addressing this challenge.
- Curation of over 200 million cells, forming the largest human multimodal dataset to date (H&E, Xenium, scRNA-seq and ATAC), which at present can only be generated at Sanger, enabling cross-tissue discovery across multiple diseases and, when used to train the right large-scale AI models, has the potential to be truly game-changing for understanding disease mechanisms and driving new discovery.
- Its commitment to the democratisation of science - success would enable democratisation of multimodal profiling, enabling predictions of genetic variation based on routinely obtained pathology samples, or predicting cell state and tissue morphology based on genetic information. As well as being widely usable nationally, this opens up opportunities to collaborate with Sanger’s existing international partners and widen access to multimodal analysis from settings with limited resources.

Grand Challenge 5: A Foundation sequence model of the Cis-Regulatory Code in developing human tissues

Science Programme: Cellular Genomics and Generative Genomics

Lead Supervisor(s): Mo Lotfollahi - Group Leader; Jussi Taipale - Senior Group Leader

The human body consists of multiple organs, each composed of millions of specialized cells that have evolved over millions of years. Despite their diversity, all cells share the same genome of 3.2 billion DNA base pairs. This cell type diversity is achieved by controlling the transcription of genes, the coding regions of the genome that form around 2% of the genome. This fine regulation of gene expression is achieved by binding a class of proteins called transcription factors (TF) to the genes and also to distal non-coding regions that are far away from the gene. The process of combinatorial TF binding-driven transcriptional regulation remains a key focus in modern biology and therefore solving this requires learning the combinatorial “grammar” of transcription factor binding and long-range regulatory interactions across hundreds of cell types within diverse chromatin contexts. Existing models such as deep generative models mostly work on gene expression space and ignore regulatory sequence, while sequence based models such as informer do not map chromatin accessibility to gene expression regulation. Recent ML/AI methods such as GET showcase the strength of using foundational models but are not yet natively multi-omic at single-cell resolution and are not tailored to deep developmental atlases (Fu et al., 2025).

We aim to advance our understanding of this mechanism through two technological breakthroughs: our ability to measure transcription in individual cells and recent advances in deep learning. The Fellow will be challenged to train a multimodal multiexpert single-cell foundation model using comprehensive in-house and public data that captures human development across various stages. This model will integrate representations of Transcription Factors and model their relationship with DNA sequences and among themselves (combinatorial binding), to predict RNA expression and DNA accessibility on a single-cell scale for the first time. We have applied this strategy in preliminary training runs using our in-house embryo data. After ~20K training steps, our model was able to predict RNA expression with a Pearson correlation score of >0.8 for all cell types, including unseen cell types. The model will provide insights into human development and fundamental principles of how our genome shapes the function of cells and their assembly into an organism.

The ‘grand challenge’ is therefore to unify these strands into a single, cell-level, multi-expert foundation model that unifies gene expression and chromatin accessibility by training on millions of cells to provide cell and chromatin context and to learn the mapping from sequence and TF context through chromatin to expression. This will be able to support zero-shot generalisation to new tissues, developmental stages and perturbations, and enables in silico perturbation of regulatory programs and disease variants.

We propose to develop a multi-expert, multimodal single-cell genomic foundation model of the cis-regulatory code, trained on multiome data that contains paired transcriptome and chromatin accessibility modalities from human development. Concretely, the model will integrate DNA sequence (including long-range context around regulatory elements), transcription factor (TF) and chromatin regulator

representations, paired single-cell chromatin accessibility (scATAC-seq) and gene expression (scRNA-seq), and relate them to rich developmental and tissue context (embryonic and fetal human organs).

Apart from building a unified model that predicts RNA expression from chromatin accessibility and the underlying genomic sequences, the goal is to map two core mechanisms that determine cell state: The cooperative interactions of sequence driven open chromatin regions to regulate gene expression, the transcription factors responsible for these interactions and finally elucidate the underlying regulatory codes that drive cell type specificity during development. The model will generalisable to unseen cell types, tissues and applicable for predicting sequence perturbations.

The challenge will be supported by Sanger's unique in-house expertise and resources:

- Human developmental and tissue atlases: Sanger has played a leading role in building multi-organ developmental atlases at scale which is mostly transcriptomic and our foundation model is poised to evolve from single modality to multiomic atlas at organism level, in this challenge, a whole embryo.
- In-house multiome and chromatin datasets: Large paired scRNA+scATAC multiome datasets from human embryos and fetal tissues provide exactly the data needed to learn cis-regulatory rules in a developmental context and enable cross tissue and co-option of these regulatory codes in diseases.
- AI and foundation-model ecosystem: The Institute is actively investing in AI to enable researchers to create high-impact technical work with broader biological reach. This challenge leverages this infrastructure to train multi-billion-parameter models to address fundamental questions in biology.
- Method novelty: This method carefully balances prior information and unbiased learning, aligning with Sanger's fundamental aim of answering the reasons behind the sequence specificity of a human genome.
- Open, reusable resources: We will release curated training datasets, model weights and inference code (as permitted by consent and governance) as community resources, matching Sanger's commitment to open science and reusable platforms.
- Applications to development, immunity and disease: The resulting model will enable prioritisation of non-coding variants, interpretation of GWAS signals in developmental and immune contexts, and in silico testing of TF perturbations, supporting Sanger's long-term goal of going from single cell to organism level.

Grand Challenge 6: AI for solving the gene regulatory code across life

Science Programme: Generative Genomics

Lead Supervisor(s): Jussi Taipale - Senior Group Leader

Modern artificial intelligence tools can solve complex recalcitrant biomedical problems that are intractable for the human mind. For example, the recent Nobel Prize in chemistry was awarded for Alphafold and Rosettafold algorithms that largely solved prediction of a protein's fold given its amino acid sequence [1]. These AI models were so successful because they were trained using large databases of experimental data, including protein structure data bank (PDB) and protein sequence data bank (GenBank). Several unsolved central problems with similar simple formulation exist, including the gene regulatory code which requires that we predict genome regulation from DNA sequence.

The gene regulatory code describes how DNA sequence determines when, where and how much genes are expressed. Despite decades of study that has revealed many of the molecular mechanisms of gene expression at a conceptual level, the problem remains largely unsolved [2,3]. What we have learned about the regulatory code revealed that it is exceptionally complex. In humans, more than 1,600 transcription factors (TFs) interact with each other to read the regulatory information in the genome. It has been estimated that 220 million parameters need to be determined to capture the effect of such TF-TF interactions on gene expression [3], and we know that more complex interactions can exist, making this a lower bound. Because different cells express different transcription factors, the code differs in cell type and organism-specific ways [3].

Some progress towards solving the gene regulatory code has been recently made using AI. Genome language models have essentially created vast genome indexes and can be used to predict sequence constraint by identifying sequences that look surprising when perturbed [4,5]. However, the regulatory genome is largely not conserved, and the regulatory code differs by organism, making these rules challenging if not impossible to learn even from many genome sequences [6]. More promising AI models built on extensive profiling of genomic sequences have also been built that appear to have a much more mechanistic understanding of the genome [7–9], but the predictive power of even the most advanced models is limited by lack of sufficiently diverse training data [3]. The vast majority of the data comes from molecular characterization of native genomes in cells. Because native genomes arose through biased evolutionary processes, learning causal relationships between DNA sequence and genome regulation is extremely challenging [3].

To solve the problem, we need to measure activities of:

1. Far larger number of regulatory DNA sequences
2. That regulate much larger set of genes
3. In most, if not all, cell types
4. For each organism of interest

Central to this plan is leveraging the power of AI, both in designing the experiments to ensure that the most informative data is generated, and in interpreting the extremely large and complex datasets generated. The final output will be a model that takes DNA sequence, cell type and species as input, and gives gene expression as output (Figure 1).

Much work is required both for designing and training the AI models, and building tools for interpretation and distillation of the models into biophysical, fully interpretable models of gene expression that can give mechanistic insight into gene regulation, and regulatory networks that control development and disease.

Solving the gene regulatory code fits exceptionally well to the strategy of Sanger and the Generative Genomics program, as it depends on large-scale data generation at Sanger, and employment of advanced AI tools. The solution to the gene regulatory code will also advance our understanding of biology, give us the ability to interpret the effects of most genetic variants that cause disease, and enable both identification of targets and development of tools for gene therapy.

References:

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
2. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. Mol. Cell 83, 373–392 (2023).
3. de Boer, C. G. & Taipale, J. Hold out the genome: a roadmap to solving the cis-regulatory code. Nature 625, 41–50 (2024).
4. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. bioRxiv 2025.02.18.638918 (2025) doi:10.1101/2025.02.18.638918.
5. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. Science 386, eado9336 (2024).
6. Jaganathan, K. et al. Predicting expression-altering promoter mutations with deep learning. Science 389, eads7373 (2025).
7. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods 18, 1196–1203 (2021).
8. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. bioRxiv 2023.08.30.555582 (2023) doi:10.1101/2023.08.30.555582.
9. Avsec, Ž. et al. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. bioRxiv 2025.06.25.661532 (2025) doi:10.1101/2025.06.25.661532.


wellcome
sanger
institute


Google DeepMind