Predicting Evolution

Towards a new ARIA programme

Yannick Wurm, Programme Director, ARIA
Richard Nichols, Professor of Evolutionary Genetics, Queen Mary University of London
Simon Evans, Science + Technology Lead - Accelerated Adaptation, ARIA

Version 0.5 from 2026-09-03

About this document

This working document presents an early draft of the core ideas underpinning a programme that is currently being explored at ARIA. We are sharing an early formulation and invite you to provide feedback to help us refine our thinking. This is not in itself a funding opportunity, but may end up leading to one.

Share your feedback and register your interest in a workshop

1. Executive Summary

"Nothing in biology makes sense except in the light of evolution." — Dobzhansky, 1973

ARIA will test whether evolution can be made forecastable. Not the fate of every mutation, and not indefinitely. But sufficiently to predict which populations will adapt, which will collapse, and how far ahead those forecasts remain reliable. Until now, our predictive successes have been largely confined to short time scales; our methods do not capture the emergent properties of biological systems that shape longer-term change. We therefore propose to initiate a step change in evolutionary forecasting by exploiting system-level patterns, which arise from the fundamental mechanisms by which genetic information is transmitted and encodes traits. Since evolution both shapes these higher level rules and is directed by them, we expect common organising principles that transcend particular species and environments.

Only now can this proposition be experimentally tested. Genome-scale measurement, data on molecular states that connect genotype to phenotype, automated evolution experiments and modern machine learning can be brought together to create—and rigorously test—forecasts at a scale previously impossible.

The bet is that evolution, despite its apparent historical contingency, contains higher-order structure that can be learnt from sufficiently rich, replicated data. In some settings we might aim for precise predictions, and in others define the range of plausible futures. At the most predictable end of this spectrum are evolutionary responses which generalise widely across the tree of life, because of the deep conservation of core cellular processes. They involve fundamental functions, including the cell replication cycle and DNA repair. In these cases, knowledge developed in model organisms such as yeast and the worm *Caenorhabditis elegans,* can be applied directly in understanding human cancer, an evolutionary span of over half a billion years, and we are able to predict the pathways and sometimes the individual genes that will respond to chemotherapy. More broadly, empirical evidence shows repeated patterns of adaptation in species separated by millions of years of evolution, albeit not necessarily the specific genes involved – rather the group of genes identified by analysis of their co-expression (e.g. in plant species ranging from weeds to trees – adapting to temperature, rainfall & seasonality). Finally, at the more abstract but possibly more pervasive level, mathematical models of the evolution of gene regulation, borrowing from statistical mechanics, suggest the nature and precision of regulatory interactions should have tractable rules across all living systems.

The radical idea that we can attain and verify evolutionary forecasting has not been empirically tested. Training and intensively testing forecasts will require experimental evolution at

unprecedented scale and replication. New predictive capability could emerge from a synthesis spanning diverse disciplines. Which are most relevant is an open question, they could include

+ our rapidly expanding mechanistic knowledge of systems biology
+ extensions of quantitative genetics
+ deep learning and other data-driven approaches

For this reason we propose a blind forecasting challenge to draw in expertise from teams from the broadest possible range of disciplines. **The winning team will provide a forecast that materially outperforms null models, retains calibrated uncertainty, and transfers to new genetic backgrounds and spatial environments.**

If this programme succeeds, its innovations may cross-inform forecasting of other complex systems (e.g. exploiting parallels between evolutionary trajectories and parameter optimization in deep learning).

Failure can also be decisive: it would establish a map of where prediction holds and where it breaks is itself field-defining, just as understanding the limits of weather forecasting or the domain of validity of Newtonian physics is as important as the predictions themselves.

2. Why is this worth doing?

Forecasting the likelihood and patterns of evolutionary change would unlock applications across agriculture, biosecurity and human health. We would like to predict when and how

+ Engineered cell lines drift from their designed function.
+ Tumours escape the treatments we prescribe.
+ Gene-edited cells delivered as therapy can fail to establish – or evolve beyond our control.
+ Agricultural pests evolve resistance to pesticides faster than we can replace them.
+ Pathogens and parasites evade drugs and vaccines.
+ Invasive species adapt and establish in new territories.
+ The release of an invasive species' natural enemies, for biocontrol, leads to successful establishment but not spread to non-target organisms.
+ Engineered crops, plants, and animal strains run amok or lose engineered properties.
+ Species shift to follow a changing environment or adapt in place.

To meet these ambitions we must move beyond our successes, limited to very short time scales, in predicting which segregating allele or new mutation will spread. Weather and climate forecasting provide a useful analogy: in the short term we can forecast the timing and character of a storm, while over longer timescales we estimate the trends in the frequency and severity. Both are useful because they support decisions – from whether to carry a waterproof today, to

whether to strengthen a roof for next year. Evolutionary forecasts could similarly inform near-term interventions, such as selective breeding or translocating individuals from particular populations to increase pest resistance or reduce inbreeding load, while longer-term forecasts should guide strategies to reduce the risk of invasive species establishing, agricultural pests escaping control, or cancers evolving resistance to treatment.

3. Programme north star

Establish evolutionary forecasting as a quantitative discipline, predicting whether, when and how a population under emerging stress adapts or collapses, precisely enough to inform the decisions that depend on it.

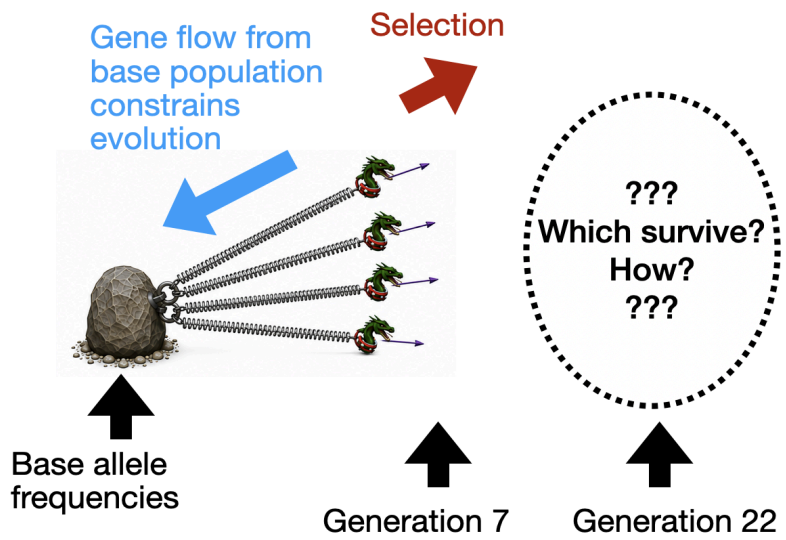
Today, evolution defies forecasting: populations diverge along paths that look chaotic. We're making the bet that massive-scale replicated evolutionary experiments, augmented with detailed molecular data and new modelling approaches can expose the structure beneath and enable the creation of new foundation models. We will test this idea through a series of open forecasting competitions, with teams making blind predictions of the experiments' outcomes, quantifying how reliably, and how far ahead, prediction holds.

4. Proposed programme structure

Overview of the experimental approach

Changes in genetic composition induced by selection can be unpredictable, reminiscent of the chaotic changes in direction of wind attributed to daemons (Amenoi Thyellai) by Greek mythology. How is it possible to get replicable and interpretable information from such erratic time courses? Our approach will be to set off multiple daemons and measure the forces projecting them in different directions, by throwing a harness around each to see how hard it pulls and in what direction, and whether it can sustain its response or dies.

How does that metaphor translate into a genetic experiment? We will apply selection to multiple populations, so that the genetic composition (the allele frequencies) begin to diverge. Our tether to constrain evolution is gene flow – every generation we will replace a percentage of the evolving population with a sample from the base population (the initial



genetic composition), diluting the genetic changes. This design has two main advantages: firstly, tethering reflects key real-world crises, where selection is localised but gene flow connects populations. Critical forecasting problems of this type include the response to the arrival of invasive species, the evolution of pesticide resistance and the response to climate change. Second, understanding the behaviour of the system becomes more tractable since we increase the signal to noise ratio by repeatedly introducing variants from the base population into the testing arena whereas, if the test populations were separated from the base, many would career off in directions determined by the variants which happened to establish in the early generations. In this schematic, the x and y directions represent changes in allele frequencies at two genomic loci. In reality there will be variants at hundreds of thousands of genetic loci, of which we expect tens of thousands to respond to selection.

Overview of challenges

Our discovery process has highlighted the need to bring in teams spanning a wide range of scientific disciplines, from evolutionary and systems biology to statistical modelling and machine learning. Each field would have distinctive assumptions and ways of approaching the forecasting problem. Our current thinking is that a prize competition could provide an effective mechanism both for recruiting and testing genuinely diverse approaches against a common challenge. Below we outline our proposed framework for the competition. A series of prize challenges of increasing ambition will drive the leap from explanation to forecasting.

Whether a population can meet the selection we impose plays out at two levels: the demographic outcome, and the genetic changes that produce it. The competitions will therefore run in two parallel tracks (demographic and genetic prediction). Teams can compete on either track, or both. The teams' forecasts can draw on genomic and biological state data we release across the first generations of selection, plus any information in our vast accumulated record of biological variation and function.

Predicting demography

For each challenge there will be a prize for forecasting demographic fate: whether existing populations persist or collapse (Challenges 1 + 2), and which habitats become populated as a species evolves to expand its range (Challenge 3).

Predicting genetic change

A companion competition will be for teams forecasting the genetic changes – after the results of the demographic competition have been released, so the identity of the surviving populations is known. In this case the teams will forecast the allele frequencies at variable sites (loci) throughout the genome.

Challenges

Challenge	Fundamental question
1.	Can models forecast adaptation and collapse of populations in a controlled system, with evolution from different genetic starting points?
2.	Do the learned rules transfer beyond the genetic backgrounds on which they were developed?
3.	Can forecasts extend to evolution in an interconnected set of sites in which alleles can spread between adjacent populations, and populations can evolve to colonise new habitats ?

Winning the prizes

In the first six months we will generate the initial data to develop the predictive tools of the different teams. This experiment will also provide data for Challenge 1.

Challenge 1

Will expose the same genetic variants to selective challenges in different combinations. We will establish base populations from 18 *C. remanei* inbred isofemale lines from a single wild source population. Different combinations will be created from a funnel crossing design, each cross excluding 3 of the lines ({1,2,3},{4,5,6} ... {16,17,18}). In this manner approximately ⅓ of the segregating sites will be entirely missing from at least one combination. After 10 generations of adaptation to the culture conditions, each line will be exposed to two forms of selection of the larval stage in liquid suspension and a control:

+ 20 replicate populations will be exposed to acute H_2O_2 stress
+ 20 replicate populations will be exposed to acute heat stress
+ 10 replicate control populations will not be exposed to any stress.

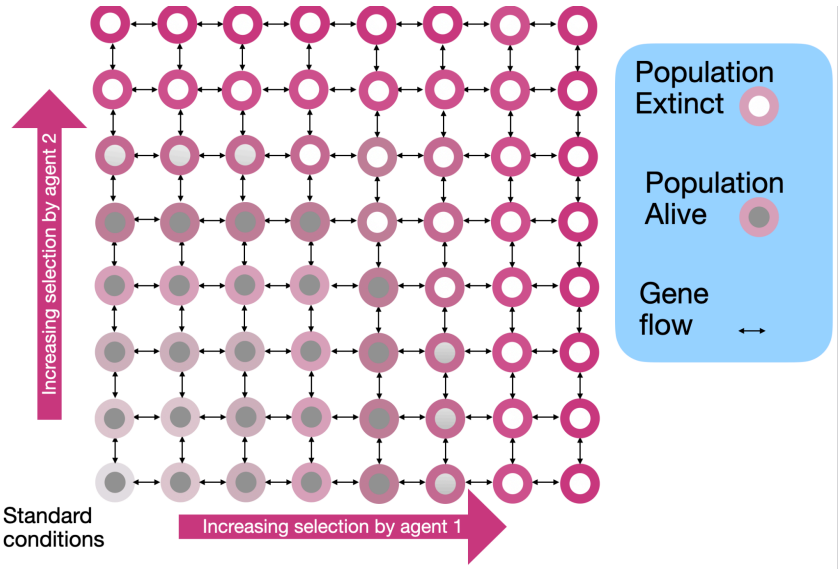
From t_0 to t_7 selection will be at an intensity which kills 40% of individuals, after which the intensity will be ratcheted up by 3.5 points per generation or until half (10) of the replicates have collapsed, i.e. no larvae hatch from the selected population. We will generate the following data from pools for each replicate and the base population: genomic sequences (9,000x for selected populations, 90,000x for the base), expression data for synchronised day-4 adults (polyA+ RNA and sncRNA) and ATAC-seq. Data will be collected at t_0 , t_4, and t_7. The data will be released after 7 generations. The demographic challenge will be to identify which populations survive. Each inbred line will be characterised by assembling the genome of 5 individuals using Picogram input multimodal sequencing (PiMmS) sequencing offered by the Tree of Life initiative at the Sanger in support of this project. After the demographic competition results have been released, teams participating in the genetic challenge will be asked to use that demographic

information (along with the BCF variant fields specifying the SNPs and alleles, including any mutations since t_7).

Challenge 2

Teams will build on challenge 1, its data and the methods it has generated. This challenge is a step up in difficulty, as it tests the teams' ability to forecast how genetic effects are reshaped in independently evolved genetic backgrounds. The lines will be created by collections from 12 new populations. They will be expanded up from large founding populations taken from new collections in the wild, without the usual inbreeding which generates blocks of correlated alleles around the genome (local linkage disequilibrium). We will assemble 100 individual genomes from each base population to characterise the haplotype structure. Line construction can be underway from the start of the project and the challenge set before the end of year 2. The timing of the challenges and the release of data will proceed as in Challenge 1.

Challenge 3



This most ambitious test will require the teams to extend their predictions to the evolutionary processes that allow a species to adapt to spread into new environments. They will forecast both the successes – adaptation to colonise habitats that were initially outside their range (empty populations in the schematic showing the start of the experiment) and the limits to evolution at the range margin. The based populations from Challenge two will be introduced to a series of populations connected by gene flow: movement of individuals from extant populations into the adjacent population in a two dimensional stepping stone scheme. Trials will be underway before the end of year 2, with the aim of completion by 3.5 years.

The selective regime in each population will be constant, and chosen to span the unselected state (bottom left) and extend beyond the limit that leads to the collapse of the population. Over time the population is expected to expand its range as evolution proceeds and, at the margins where expansion fails, we expect to see loss of genetic diversity, a change in the distribution of the effects of surviving alleles, selection for new genetic combinations (changes to linkage disequilibrium and genomic architecture) and regulatory interactions. If innovations in automation allow, the scheme will be replicated 25 times. The experiment will be run for 80 generations. The teams will be invited to predict which populations produce offspring that

survive the selective challenge (a list of populations for each generation in a 20-generation time course), and the distribution of each allele among the populations. Competitions will be run at 20 generation intervals after the release of data at t_{20}, t_{40}, and t_{60}.

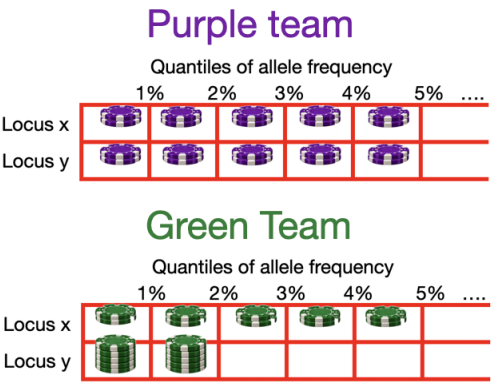
Scoring

Demographic scoring

For challenges 1 + 2: Once the experiment is run, we will release the proportion, *f*, of the 20 replicates that persisted (produce surviving larvae before gene flow) in each of the lines for each selection regime. The teams' goal is to forecast which those are. The forecast will be returned as the persistence probability, p_i, for each replicate. Their score will be log(p_i / *f*) for populations that persist and log(1-p_i / 1-*f*) for those that fail. The sum will be zero if they simply predict *f* for every population, and will be larger the more accurate their p_i values are. For challenge 3 the value of *f* will only be released for the first generation of the 20 generation series.

Genetic scoring

For challenges 1 + 2: Where selection is applied to single populations, the forecasting procedure can be thought of as placing tokens on a board in a game of chance. The squares represent a range of allele frequencies which are equiprobably under random genetic drift. The team's score is the number of tokens staked on the correct outcome (they have 100 in total per locus, and can stake a proportion of a token). In these illustrative examples the Purple Team forecasts that allele divergence is shaped by random drift rather than by selection - hence they spread their bets evenly. The Green Team agrees that locus x is shaped by genetic drift, but predicts selection will act strongly on locus y (an intense selective sweep). If the reality for the population is that the alternative allele is being driven towards fixation, the Green team will therefore beat the Purple team (but Purple wins otherwise).

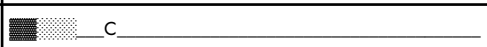
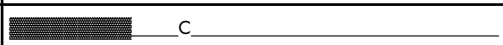
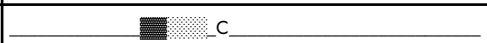
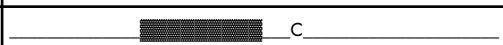
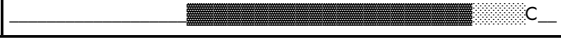


For challenge 3:
Where selection is applied to a system of interconnected populations the teams will be asked to specify **where** each allele is found. They will do this by forecasting the relative occurrence of the allele in each population, f_i ($\sum_i f_i = 1$), so the expected frequency of an allele with mean frequency $\bar{p}$ over *n* populations is $p_i = n\bar{p}f_i$. The score rewards the team that matches the

observed frequency, *o,* with the greatest accuracy: - $\sum_i (p_i - o_i)^2 / o_i(1-o_i)$. The values of $\bar{p}$ will be released along with the results of the demographic competition.

Advanced Research + Invention Agency ARIA

Overview of competition timelines

Competition	Track	Closes	Experiment	Data Released at Open	Prediction Target	Timeline (M1 – M48) ▓: release of genetic data; ▓ density data; ▒: more evolution; C: competition
C1	D	M9	6 intercrossed lines. 2 stressors + control (20 + 20 + 10 populations per line)	Genomes and expression at t0, t4, t7	Persistence probability per replicate at t22	
C1	G	M15	Same	All C1 data; demographic outcomes; new mutations t7 – t22	Allele frequencies at t22	
C2	D	M21	12 wild-derived lines. 2 stressors + control (20 + 20 + 10 populations per line)	Genomes and expression at t0, t4, t7	Persistence probability per replicate at t22	
C2	G	M27	Same	All C2 data; demographic outcomes; circular-cross t22 frequencies	Allele frequencies at t22; revised intercross forecast	
C3.1	D + G	M33	Two-dimensional stepping-stone matrix; 12 lines; 80 generations	Lattice occupancy and genomes, t0 – t20	Occupancy and allele distribution at t20	
C3.2	D + G	M39	Matrix continued	All lattice data to t40	Occupancy and allele distribution at t60	
C3.3	D + G	M45	Matrix continued	All lattice data to t60	Occupancy and allele distribution at t80	

5. Why this experimental design?

Why selection with gene flow?

The gene flow from very large base population (an effective size over half a million) has three main benefits

- It makes loss of genetic diversity minimal, allowing the ongoing resupply of ancestral alleles to the selected population by gene flow. Crucially, the hundreds of thousands rounds of sexual reproduction will generate new **combinations** of alleles every generation, so alleles are introduced into the selected population on a different genetic background and tested.
- This extensive recombination **breaks down linkage** between selected alleles and adjacent loci on the genome, so that hitchhiking variants are less easily mistaken for the actual targets of selection.
- The gene flow from a large and stable reservoir restraints the response to selection so the allele-frequency divergence from the source provides a **quantitative readout of the strength of selection**, and how it varies from place to place around the genome.

Why use robotics and automation to ensure many large selected populations?

- The large size increases the signal to noise ratio, since the effect of random drift on the equilibrium frequency is minimised allowing the detection of small frequency differences and hence **the fine scale response to selection**.
- Combined with the unprecedented scale of replication provides a degree of **precision in the selection estimates** exceeding that which has accounted for the majority of additive genetic variance in other systems.

Why a multicellular eukaryote?

Experimental evolution on a eukaryote with well characterised tissues and development has important advantages. Extensive studies in developmental and functional genomics have revealed systematic differences in genes' evolution according to when and where they act: genes central to development or deployed across many tissues, or many environments tend to be more constrained, while more restricted functions provide greater scope for evolutionary change. The teams can exploit this rich prior knowledge—developmental timing, tissue specificity, regulatory interactions and gene function—to ask whether the observed patterns of evolutionary change reflect a learnable architecture of constraint and innovation. Because these features of development are shared across multicellular organisms, the resulting representations offer a route to

extrapolating beyond the experimental species.

Why *Caenorhabditis remanei*?

It has the advantage of very short generation time ~ 4 days. It has two sexes and outcrossing, hence sexual reproduction and recombination is ensured each generation, creating new combinations, and including the mixing of selected and unselected lineages. It is closely related to *C. elegans* (selfing species) for which we have near-complete cell-lineage and connectome maps, extensive functional annotation, mutant collections and large-scale expression and regulatory datasets. Its close relationship to C. remanei allows these resources to help interpret molecular states and genetic changes in the experimental system without sacrificing the abundant standing variation and obligate outcrossing needed for evolutionary forecasting.

Why 20 replicates of each line and selection regime?

Our target was to achieve the level of precision in identifying loci that contribute to quantitative traits required to explain their heritability (to solve the missing-heritability problem – effects of 0.005x additive genetic standard deviation). Using conventional automation and rearing (9cm Petri dishes) we can achieve two selection regimes with 20x replication, with 10x replication of controls, and a 10x base population in an experimental block of 3600 dishes at an effective population size of 60,000 per replicate (60 dishes), which achieves 80% power to detect such small effects.

How can we learn about millions of loci?

At first sight, learning to forecast evolution from trajectories at one million loci across 1,600 replicate populations appears hopelessly underdetermined. In reality, the problem is strongly constrained by multiple independent sources of biological knowledge. Under simple additive models, selected loci follow characteristic allele-frequency trajectories whose expected form is well understood, while the much larger set of neutral loci provides an empirical estimate of the stochastic background generated by drift, migration and sampling - which parameterise the noise affecting the trajectories of selected loci. Independent replicate populations repeatedly place alleles into different genetic backgrounds, allowing epistatic interactions to be inferred. Prior knowledge from C. elegans further constrains the search space through known regulatory and protein interaction networks, while transcriptomic and epigenetic measurements provide intermediate molecular states needed to draw inferences on the genotype-phenotype map. Finally, adaptation versus collapse supplies a powerful supervised learning signal linking these genomic and molecular trajectories to evolutionary outcome.

This task is analogous to modern deep learning for computer vision. An image-recognition network is not expected to infer the concept of a face directly from millions of pixel intensities. Instead, it progressively learns a hierarchy of increasingly informative representations—edges, textures, object parts and complete objects—before reaching the final classification. The intermediate representations dramatically simplify the inference problem. Likewise, evolutionary forecasting can construct intermediate representations of the organism's latent state, allowing complex patterns of genetic interaction to emerge naturally. Rather than learning millions of independent locus effects and their interactions, it learns the lower-dimensional biological processes that govern adaptation and collapse.

6. Implementation

The experimental design requires a tractable laboratory organism with a short generation time. A eukaryote with a well understood development and well defined tissues will reveal patterns in the rate of genetic innovation as a function of tissue and expression timing of different loci, which will allow extrapolation across the multicellular branches of the tree of life. Having two sexes and obligatory sexual reproduction means that selection with gene flow is implemented, reflecting real-world challenges in which severe selection is localised in part of the species range.

We will initially use 18 existing isofemale lines of *Caenorhabditis remanei* that have been inbred for 10 generations and capture the genetic variation from a single population. Challenges 2 and 3 will use new lines specially collected for this programme from wild populations to exploit the high genetic diversity, and different haplotype structure. With traditional rearing we can maintain our desired base population size of over N_e = 500,000 with eight 1m stacks of 9 Cm plates (each N_e = 1,000), and have 20 replicates and 10 controls (each of one stack) selected by two selective agents (e.g. the temperature 36°C+ for 40 mins, rising to 37°C+ as selection responds, or H2O2 exposure for 40mins 0.75 - 2 nM) to maintain 40% mortality.

The large base population is essential to ensure that the genetic diversity is maintained and that new combinations generated by recombination are being introduced to the evolutionary mix. The size of each selected population will be N_e = 60,000 - which reduces genetic drift sufficiently to detect small effect sizes. Assuming an additive model the 20x replication would allow the identification of effect sizes of less than 1% of the standard deviation of the additive variance after 22 generations, a precision which has proved sufficient to essentially close the 'missing heritability' gap (in the very largest human GWAS studies).

7. What we're still trying to figure out

- If this fails, how or why does it fail?
- Could us doing this differently work better or be more impactful?
- How can the competition format be improved? Be made more resilient? Is every 6 months the right cadence?
- How can we enable small creative teams (from academics to startups) to best participate? What should be the shape of support we provide?
- Which other organisations might complement our investment to enhance the competition? Might there be organisations willing to:
    - contribute compute credits?
    - increase the prize pots or give stipends to creative teams with high potential?
    - store data in their cloud infrastructure?
- What should competitors supply - simply their predictions? Or should they also be forced to share their algorithms or learning/protocols (and if so, must those be shared immediately? Or after a certain timeline?). Or should "closed" competitors be in a different track or ineligible for prizes.
- How can the risks of cheating be minimised?
- Which data formats should datasets be shared in?

Share your feedback and register for a workshop to discuss this proposed programme

8. Approximate budget

This programme is still being explored and is not a funding opportunity, but it may end up leading to one. This budget is intended as a prompt for feedback on how resources should be allocated for a programme exploring this area, and is not indicative of actual planned spending.

		Y1	Y2	Y3	Y4
TA1: Running the competion	£2,000,000				
TA2: Experimental evolution					
TA1.1 run experiments manually (year 1)					
+ running the experiments	£5,000,000				
+ generating & storing sequence data	£5,000,000				
TA1.2 use robots (years 2-4)					
+ running the experiments	£10,000,000				
+ generating & storing sequence data	£20,000,000				
TA3: Tooling & Automation					
TA2.1: Build the robots to automate experimental handling	£4,000,000				
Requires local temperature control of each experimental unit					
TA2.2: Build 3d structures with 100x the population carrying capacity of petri dishes	£1,000,000				
TA4: Prime new modelling capability					
Giving creative small teams pilot funds of 150k	£4,000,000				
e.g., by providing compute resources or credits	£4,000,000				
TA5: Competition prizes					
10 competitions * 750k prize pot / competition.					
C1 part 1 + 2: grand prizes of 50k. Other prizes of 10-25k	2000000				
C2 part 1 + 2: grand prizes of 100k. Other prizes of 25-50k	2000000				
C3.1 part 1 + 2: grand prizes of 250k. Other prizes of 75-100k	3000000				
C 3.2 part 1 + 2: grand prizes of 500k. Other prizes of 100-250k	3000000				
C 3.3 part 1 + 2: grand prizes of 500k. Other prizes of 100-250k	3000000				
Total	£68,000,000				