# Programme Thesis

# Forecasting Evolution

v2

Yannick Wurm, Programme Director

Simon Evans, Science + Technology Lead

Richard Nichols, Science Contractor

## CONTEXT

This document presents the core thesis underpinning a programme that is currently in development at ARIA. We share an early formulation and invite you to provide feedback to help us refine our thinking.

This is not a funding opportunity, but in most cases will lead to one – sign up here to learn about any funding opportunities derived or adapted from this programme formulation.

An ARIA programme seeks to unlock a scientific or technical capability that

+ changes the perception of what's possible or valuable
+ has the potential to catalyse massive social and economic returns
+ is unlikely to be achieved without ARIA's intervention.

## THESIS, SIMPLY STATED

*Nothing in biology makes sense except in the light of evolution.*
Theodosius Dobzhansky, 1973

Dobzhansky was right, but the light he describes illuminates only the past. Evolutionary biology can explain in exquisite detail why a population became what it is; it struggles to say what will happen next. This means management decisions are made in the dark: *will a population adapt fast enough to persist?*

Sometimes we want evolution to succeed; other times we need it to fail. For example:

+ Will a herbicide eradicate a weed from a farm, or leave a resistant population to spread?
+ Will an engineered organism retain its designed function?
+ Which agricultural animals and plants will be able to keep pace with climate change?
+ Will a new disease wipe out a wild population, or can resistance evolve fast enough to save it?
+ Will a course of treatment eradicate a tumour or clear the way for treatment-resistant cells to grow?

**Imagine if we could forecast evolutionary outcomes early enough to act. Evolution would stop being a force we react to and become one we work with. That would transform how we treat disease, grow food, protect wild populations and engineer living systems.**

In every example, inherited differences affect survival and reproduction, changing the population over generations. The question is whether that change happens fast enough to determine the outcome.

Evolution determines whether key decisions in sectors worth trillions of pounds deliver the outcomes we intend. Yet we cannot reliably forecast it. Answers require the equivalent of what meteorology has built for weather: reliable statements of what will happen, when, and with quantified uncertainty.

We are thus setting the following forecasting challenge:

***Across thousands of experimental populations facing increasingly hostile conditions, and given their population counts, genetic and gene activity information: which populations will adapt fast enough to survive, which will collapse, and when?***

Except for rare, simple scenarios, answers are elusive. Our bet is that evolutionary responses contain recurring patterns that new modelling approaches can learn from rich, replicated

data. The challenge is to identify those patterns amid the noise and turn them into reliable forecasts of population fate.

Three technological developments make this programme possible for the first time:

+ **DNA sequencing and molecular profiling** are now cheap enough to measure genetic composition and gene activity levels repeatedly, across large numbers of individuals and populations.
+ **Commoditised robotics and AI** enable us to perform controlled evolution experiments at unprecedented scale.
+ **New machine learning** approaches can find structure in datasets too large and noisy for conventional analysis.

We will harness the first two to build what a forecast must be tested against: a large, standardised, open record of what evolution actually did, under known conditions. Evolution has no dataset equivalent to the Protein Data Bank, ImageNet, or weather observations. We will build the matched genetic and population trajectories that are needed through massively replicated and measured experimental evolution in a nematode worm: a short-lived, sexual, recombining, genetically diverse animal, whose populations can be reared and controlled at scale.

Teams will compete to forecast population size and genetic change under heat and chemical stress. Open competition draws in approaches from outside evolutionary biology, and lets us concentrate funding on the data that forecasting requires.

After a six-month feasibility and calibration stage, challenges progress from a single habitat to connected habitats and, if forecasting performance supports it, to real-world systems. Pre-set gates stop the programme if feasibility or progress falls short.

Dobzhansky's light shines only on the past because nobody has yet built the instrument that swings it round to illuminate the future. That is what we propose to build: a discipline dedicated to *forecasting evolution*. Our programme will reveal how far into the future it can see.

# THESIS, EXPLAINED

*A detailed description of the programme thesis, presented for constructive feedback*

## Why this programme?

Evolution is not a historical curiosity. It is happening all around us: in every pathogen evolving resistance to a drug, every pest escaping biocidal control, every wild population struggling to deal with climate change. In each case, the capability gap between anticipating that change and only documenting it afterwards is what separates a successful intervention from one that comes too late.

Unfortunately, evolutionary biology remains overwhelmingly an explanatory science, rather than a predictive one. It can account for why a population evolved as it did after the fact, often in exquisite detail. But breeding decisions, drug approvals, and biocontrol releases all have to be made before the evolutionary outcome is known. The missing capability is to turn that explanatory understanding into reliable forecasts before evolutionary outcomes are known.

Approaching evolution as a forecasting problem is a distinct challenge for machine learning, not only for biology. Existing biological foundation models (e.g., AlphaFold[1], ESM[2], Evo[3]) learnt structural and functional regularities from vast databases of static snapshots. Forecasting evolution requires something categorically different: models that learn how those regularities change over time under selection, and that generalise not by memorising one population's history but by capturing rules that generalise to new populations and conditions, given each population's genetic background and the pressures acting on it. Testing whether such models are possible using data at a scale no existing dataset provides is as much a contribution to machine learning - where financial time series forecasting is the hidden centre of innovation - as it is to evolutionary biology.

## *What we hope to accomplish*

We aim to trigger the creation of a new evolutionary forecasting industry.

**Programme North Star: Forecast whether and when populations will adapt fast enough to survive, early enough to act.**

Forecasting population size through time is how we forecast whether a population will persist or become extinct, and when extinction may occur.

From census, sequence and expression data collected up to a defined forecasting point (e.g., the first seven generations for Challenge A), competitors forecast population sizes at each subsequent generation, whether and when populations collapse or colonise new habitats, and how gene variant frequencies shift. Difficulty ramps from heat or chemical stress in single habitat (Challenge A), to a spatial grid (Challenge B) to real-world systems (Challenge C). Forecast windows and inputs are specified for each challenge; success criteria are set out under "Gates" and in the Annex.

For each Challenge C test, partners will specify the practical decision and minimum useful forecast lead time before forecasts are submitted.

Success is not one correct prediction. It is the delivery of forecasts that a decision-maker can act on: a range of outcomes with probabilities attached, and an honest statement of how far ahead the forecast holds. Weather forecasting met that standard not with one model of indefinite precision but with a hierarchy of models, each suited to a different horizon, each explicit about where its skill runs out. A forecast is useful not because it predicts everything, but because it states what can be predicted, how reliably and for how long. The decisions such forecasts inform range from near-term interventions, such as which drug to give or which variants to breed for, to longer-term strategy, such as whether an invasive species will establish or a pest will escape control[4].

We do not yet know whether evolutionary trajectories are predictable in principle, only that they have rarely been predicted in practice. Our bet is that evolutionary trajectories only look disordered. They are historically contingent and seemingly one-off, yet there is signal amidst the noise, encoded in molecular and genomic information that can now be measured cheaply and repeatedly. Forecasting evolution means identifying that structure and interpreting it.

We expect that structure to take the form of recurring regularities: what appears to be an enormous genome-wide space of evolutionary possibilities may resolve into a small set of repeatedly accessible, functionally related genetic and regulatory responses to natural

selection. If so, a model need not track millions of loci independently: it needs to find the far smaller number of coordinated patterns that constrain what happens next.

There is precedent for exploiting such regularities in GraphCast, which beat the leading operational deterministic weather forecasting system on 90% of forecasting targets[5]. GraphCast does not forecast each grid cell of the atmosphere independently; rather, its graph representation learns coupled spatial and temporal structure, presumably the fronts, jets, storms and large-scale circulation, atmospheric processes that jointly constrain future states.

The step we propose to test is whether genomic, molecular, phenotypic and demographic measurements show such regularities, and whether models built on them can forecast not only which variants rise but whether adaptation is fast enough to prevent collapse.

The modelling challenge is to combine genomic information with short, correlated time series and known environmental pressures, and test whether the resulting forecasts generalise across populations, environments and biological systems. Evidence from financial modelling and weather/climate forecasting suggests promising approaches for learning predictive structure from noisy, correlated data.

## Why ARIA?

This needs ARIA because the economics are the wrong shape for anyone else. The benefits are diffuse, long-term and shared across many fields and industries. But the costs of experimentation and sequencing at unprecedented scale are concentrated upfront and carry a real chance of failure. No laboratory, company or conventional funder has the remit or the incentive to carry that cost. And none currently does: the evolution-forecasting work led by DARPA (Defense Advanced Research Projects Agency) targeted viruses, while ERC (European Research Council) and NSF (National Science Foundation) grants on animals or plants support experiments with tens of small populations, not thousands of large ones.

Here, ARIA funds the data generation; we're setting up competition for most of the rest. Open data allow teams to compete using their own resources (supplemented by targeted development awards); competitors retain the models, talent and visibility they build. Open competitions also draw solutions from outside the founding field, as AlphaFold did at CASP (Critical Assessment of Structure Prediction).

Finally, the UK can lead. We have world-class tradition and depth of expertise in quantitative genetics, breeding, experimental evolution, genomics and machine learning, and a vibrant ecosystem of both small and large companies innovating in this space.

## Why it's worth shooting for

Applications differ in two key ways: how much data exist, and how far ahead a forecast must reach to influence decisions. In most, current practice detects change after it has happened or tracks the present state; neither tells you what comes next. The lead times below are the time horizons that decisions typically require.

| Application | The decision a forecast would change | Lead time | Current data |
|---|---|---|---|
| **Biocide resistance** in weeds, pests and livestock parasites | Agrochemical and animal-health companies setting product lifespans, stewardship and combination strategies; regulators and resistance-management bodies specifying treatment rotation and dosing | 1–5 seasons | Partial. Assays detect resistance once present; historical tracking is rare |
| **Tumour escape** | Clinicians designing therapeutic regimens; companies designing trials; ctDNA (i.e., sequencing tumour DNA from patients' bloodstreams) becoming forecasting inputs rather than relapse alarms | Weeks to months | Substantial. Serial ctDNA gives repeated allele frequencies under known treatment |
| **Agricultural breeding** | Crop and livestock breeders choosing what to select for against a moving climate target | 1–10 generations | Strong pedigree and phenotype records, genomic records uneven |
| **Invasive species and biocontrol** | Biosecurity agencies setting border and surveillance priorities; bodies licensing biocontrol releases; managers choosing eradication versus containment | 1–5 years for insects; decades for long-lived species | Poor. Records are presence and absence; genomic time series are opportunistic |
| **Engineered organisms** | Manufacturers deciding when to replace a production culture; regulators specifying genetic-stability evidence; developers of gene drives and engineered biocontrol, for whom containment is a gating question | Tens to hundreds of generations | Laboratory stability measurements exist; availability varies by organism and setting |

| | | | |
|---|---|---|---|
| **Wild populations** | Statutory conservation bodies deciding where intervention is cost-effective; restoration planners choosing provenances | 1–10 generations: years for an annual plant or short-lived animal, decades for trees and long-lived vertebrates | Poor but improving. Darwin Tree of Life supplies reference genomes; museum collections supply occasional time series |
| **Harvested wild populations** | Stock assessment bodies setting harvest selectivity | Decadal | Archived tissue provides material for genomic time series |

## From nematode to impact

This programme's experimental work will largely focus on a nematode worm (see "Why a nematode?" section below). The programme's impact comes through three general directions:

+ **A proof point that changes what is considered possible.** Demonstrating multi-generational forecasting in our controlled setup will be the evidence that prompts others to follow suit.
+ **General rules, modelling approaches, model parameterisations, and mechanisms.** Biological mechanisms are rarely species-specific; the same goes for the measurements or the modelling approaches. For some contexts (e.g., livestock-infecting nematodes) only minimal adjustments may be required.
+ **Transfer/successor challenges C**. Models will need to be adapted and tested for each application. Challenges focused on transfer are built into the later stages of the programme, where we will partner with others to generate genomic time-series that can underpin domain-specific forecasting challenges

There are strong grounds to expect some cross-species transferability of forecasting capability. Population genetics theory applies across species, from maize to cattle to humans. The selective pressures we're using, heat and oxidative stress, act through molecular pathways that are conserved from worms to humans. Shared molecular functions can persist despite extensive divergence: half of genes that are essential to yeast can be replaced by human versions despite 1 billion years of divergence[6]. Distantly related plants adapt to climate through changes involving the same gene groups[7]. Gene variant effect predictors trained from species across the tree of life predict which human mutations cause disease[8]. Evolutionary forecasts have also transferred between related bacterial species, with greater consistency in predicted phenotypes than in the precise genetic routes producing them[9]. These examples suggest that shared biological functions, processes, and constraints

could provide reusable information for forecasting evolutionary change. We will explicitly test this proposition through Challenge C.

Successful transfer would give users in the tested application areas a new basis for acting before resistance, collapse or escape. Partial transfer would still identify which measurements and modelling approaches carry over, what needs further development, and whether forecasts provide enough time to act. Even an incomplete map of these capabilities would help researchers, companies and funders decide which applications to pursue next, and what evidence and investment they require.

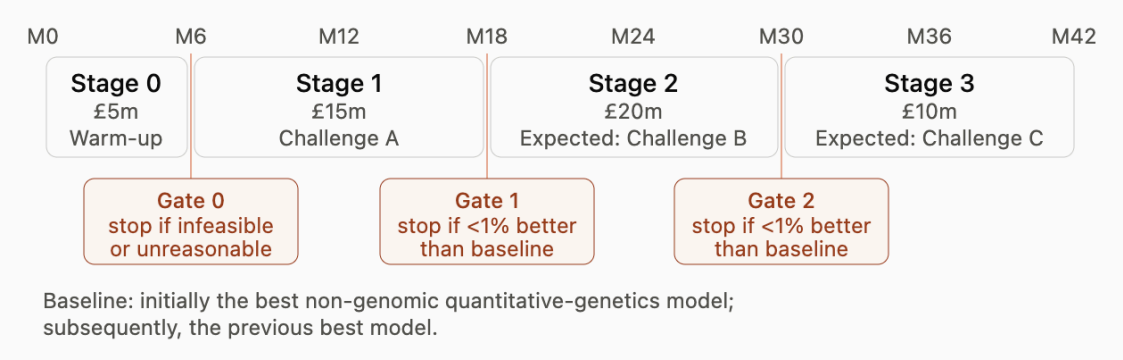
## How the experiment and the competition work

The proposed programme has five technical areas (details in "What we expect to fund"), each delivered by one or more teams or partners:

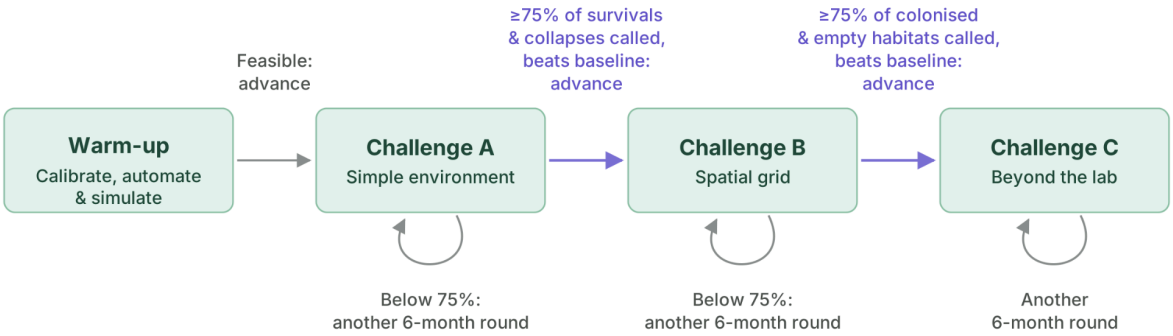
+ **TA0, feasibility**: test feasibility, calibrate experiments, trial automation, and run simulations.
+ **TA1, experiments:** run evolution experiments and record outcomes
+ **TA2, infrastructure and assessment**: release data, run the competition, hold outcomes back, and score forecasts.
+ **TA3, forecasting**: use the data to forecast population size and genetic change.
+ **TA4, transfer partners** provide real-world systems for Challenge C.

Our Programme consists of evolutionary forecasting challenges of escalating difficulty during which experimental teams generate the data used by competing teams to generate their forecasts.

We do not know how good forecasts may be. The data, models, or competitors may fall short of what forecasting requires. Forecasts could also exceed our expectations. The programme will be structured in four stages. At three gates, the programme stops unless it meets pre-set criteria: feasibility at month 6 (Gate 0), and forecasting improvement at months 18 and 30 (respectively gates 1 and 2).



The programme will have an initial 6 month feasibility and calibration stage 0. Subsequently, each stage will include two 6-month challenge rounds. Forecasting accuracy decides which challenge runs next. Successive rounds of Challenge A use different starting genetic backgrounds. Challenge B may instead follow continuing populations through successive forecasting rounds.




The subsections below provide additional context for organism choice, how a challenge runs, and how the challenges escalate; additional technical details are provided in the Annex.

## Why a nematode

Forecasting long-term evolutionary outcomes in the contexts mentioned above is too costly to attempt without evidence it could work. Our programme thus aims to demonstrate feasibility in a system that is simpler… but not too simple.

Many experiments with bacteria or viruses are effectively clonal; our design instead recombines standing genetic variation every generation. Unicellular eukaryotes such as yeast or single-celled algae have more complex molecular machinery, but the evolution of such organisms is less likely to be constrained by the tradeoffs faced by organisms whose genes act across multiple tissues and developmental stages. Such constraints may make evolution more predictable. Among traditional laboratory organisms, generation times of Arabidopsis thaliana thale cress is 6-8 weeks, fruit flies (generation time of 10 days) are challenging subjects for automated handling and trait measurement, and the classic nematode Caenorhabditis elegans mostly self-fertilises and has exceptionally low genetic diversity.

| System | Sexual reproduction with outcrossing and recombination every generation | Standing genetic diversity | Tissues and development | Generation time | Robot handling at scale |
|---|---|---|---|---|---|
| Viruses, bacteria | ✗ | ~ | ✗ | minutes–hours | ✓ |
| Yeast, single-celled algae | ~ | ~ | ✗ | hours | ✓ |

| System | Sexual reproduction with outcrossing and recombination every generation | Standing genetic diversity | Tissues and development | Generation time | Robot handling at scale |
|---|---|---|---|---|---|
| Thale cress (*Arabidopsis thaliana*) | ✗ | ~ | ✓ | 6–8 weeks | ✗ |
| Fruit fly (*Drosophila melanogaster*) | ✓ | ✓ | ✓ | ~10 days | ~ |
| *Caenorhabditis elegans* | ✗ | ✗ | ✓ | 3–4 days | ✓ |
| ***Caenorhabditis remanei*** | ✓ | ✓ | ✓ | 3–4 days | ✓ |

*C. remanei* is likely the most complex animal in which controlled condition experimentation at the scale needed for forecasting is feasible.

+ C. *remanei* is genetically diverse and males and females must mate, such that this diversity is recombined in every generation.
+ Generation time: 3-4 days → 20 generations within three months
+ Large population sizes can be scalably managed using laboratory robots: 10,000 can live on standard laboratory plates; worms can be managed, counted, sorted and dispensed using microfluidics.
+ *C. remanei* shares common ancestry with *C. elegans* 20-60 million years ago[10]. Interpretation work can thus inherit the toolkit of the best characterised animals: characterised cell lineage, connectome, functional annotation, mutant collections, and large expression and regulatory datasets.

## How a challenge runs

The nematode-focused challenges will follow generally the same steps. We describe the broad concepts here for the first round of Challenge A.

### 1. Starting populations

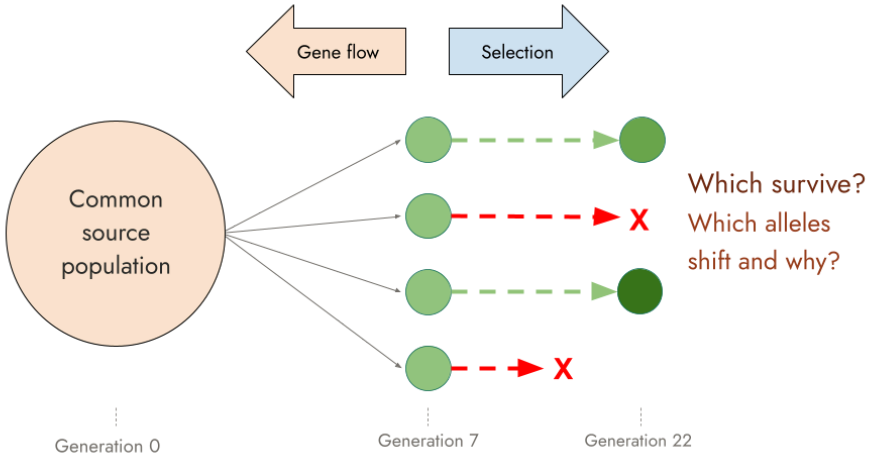
We will create ten large starting populations by collecting worms from wild populations; a population will be grown up from 1000 worms, each taken from a different host isopod (a woodlouse). Each population will be grown up rapidly to a census size of over 1,000,000 individuals (500 plates; estimated effective population size 500,000) and maintained in culture conditions for 10 generations before selection experiments begin. Each population will carry genetic polymorphism at over five million single nucleotide sites. Even in the chromosome centres, where recombination is relatively infrequent, the association between neighbouring genetic variants typically halves over only about 500 base pairs, encompassing around 20 variable sites.

## 2. Increasing stress:

Each starting population is used to found a base population and 50 experimental populations: 10 unstressed controls, 20 we will expose to acute heat stress, and 20 we will expose to acute oxidative stress ($H_2O_2$). This replication is because we want to explore the multiple solution pathways that can exist for a given environmental challenge. Each round of Challenge A therefore comprises 500 populations, 400 of them under stress.
At each generation, the stressor will be applied to L1 larvae for four hours. It will initially be set to LD40 (the intensity that kills 40% of individuals across the experiment at that generation). That intensity will increase because LD40 will be recalculated every generation. From the 7th generation, the intensity will be increased at the rate determined by the warm up package, with the aim of approximately 50% of stressed populations collapsing by generation 22.

## 3. Migration from source population for realism and increased signal-to-noise:

The genetics of isolated populations can drift apart due to stochasticity. Here, every generation, we will introduce 10% of individuals from the original starting population, provided the experimental population produced surviving larvae in the preceding generation. If it produced none, it is extinct and receives no further migrants. This has four impacts:



Continuing gene flow from the base population constrains how far each replicate can diverge

+ It *mirrors real-world problems* where selective pressures are local, but new individuals keep migrating from elsewhere[11,12].
+ Thanks to recombination, it separates the targets of natural selection from neighbouring "hitchhiking" genetic variants, thus increasing signal-to-noise ratio.
+ It restrains the response to selection, so the change in frequency of a gene variant in a population provides a direct measure of the strength of selection on that variant.

+ It ensures that genetic variants that could be lost due to stochasticity are tested again on new genetic backgrounds.

4. **Measure what's happening:**

   We generate genome sequences and gene activity measurement (mRNA sequencing) from pools of individuals from each replicate at generations 0, 4, 7, 12, 17, and the final generation (generation 22 or the last generation before collapse). To understand how genetic variants are coinherited within chromosomes, we generate full individual genomes from 1,000 individuals per starting population. We count the number of larvae produced by each population in each generation, before gene flow. Each generation is archived, allowing later analysis of the final generation before collapse.

   Only the data from generations 0 to 7 are released to competitors prior to forecasting. Released inputs include stressor identity, intensity and exposure duration through generation 7, together with the protocol governing subsequent stress escalation and migration.

5. **Competitors produce forecasts for one or both competition tracks**

   **Demographic track**: Competitor teams forecast the size of each population at each generation. We ask for probability distributions, because they let teams express how confident they are. Teams may instead submit single numbers, which are scored as forecasts made with certainty.

   **Genetic track:** Once demographic outcomes are released, and competitors know which populations survived, they can choose to also forecast gene variant frequencies at generation 22 or the last generation before collapse.

   The assessor team (TA2) scores both tracks. Each round has its own prize pot, and multiple prizes can be awarded. Submission requirements, including the treatment of incomplete or late entries, are published by the competition administrator (TA2) before each round opens.

## Stages, gates and challenges with escalating difficulty

Gates at months 6, 18 and 30 decide whether the programme continues. Progression to a more complex challenge is assessed after each six-month round, but requires at least two completed rounds of the current challenge. Stopping at any gate still delivers an open dataset and a map of where forecasting fails.

### Warm up stage 0 (6 months): calibrating experiments & technical feasibility

This stage essentially aims to set up logistics, identify challenges, and understand feasibility for running stage 1. It will also identify contexts for improving the experimental design.

We will run small- to medium-scale animal experiments, which are possible using conventional rearing (manual manipulation of plates). In order to start immediately we will use established laboratory lines. Available lines have genetic diversity up to 60% of those we propose to use in the main experiment, but with coarser genetic localisation, because inherited variants remain linked across megabase-scale DNA blocks. These resources will allow proof of principle for key elements of the experimental design, and to answer specific questions:

+ How fast must stress rise to collapse populations?
+ Do inbred lines collapse sooner than genetically diverse ones (i.e., does genetics matter?)
+ How often replicates of a population meet different fates, or meet similar fates through different molecular/evolutionary pathways.
+ Do replicates only follow different molecular/evolutionary pathways if they have different genetic compositions at the start?

We will bring new wild *C. remanei* into lab culture for Challenge A.

We will scale up rearing and automation techniques. This is essential to meet the ambitions of scale our experiments require.

Stage 0 is delivered by a coordinated set of warm-up teams (TA0) from several disciplines. Their work informs who we contract for Stage 1, and TA0 teams may bid for later technical areas.

We will perform full-genome individual-based simulations with two aims:

+ Providing a stronger evidence base for Challenge A+B experimental designs than we currently have based on analytical techniques
+ Releasing datasets so teams can prepare. We will run a test challenge on simulated data (no prizes).

### Challenge A: is evolution forecastable with a simple environment

Subject to programme continuation at the gates, we will run Challenge A (see "How a challenge runs") for at least two six-month rounds.

Challenge A involves simple scenarios where a single parameter (heat or chemical stress) increases over time. Despite the simplicity, we anticipate modest results from

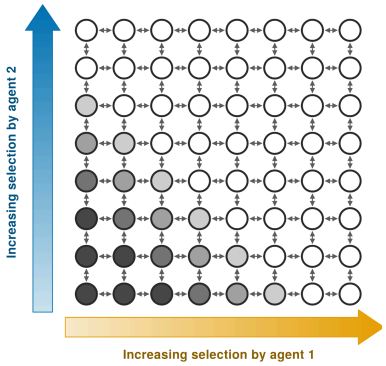
the first round of challenge A, as competitors will have almost no longitudinal training data.

After demographic submissions close, population size data from every generation and every population are released. Genetic submissions are then made with demographic outcomes known, while later-generation genomic and gene expression data remain withheld. Those data, from generations 12, 17 and 22 (or the last generation before collapse), are released only after genetic submissions close. The completed releases provide training data for subsequent rounds.

Teams will use early-generation data and outcomes from previous rounds to forecast populations with different genetic starting points. The aim is for forecasting success to depend on capturing the deeper dynamics governing evolutionary trajectories, rather than extrapolating short-term trends.

During Stage 1, we will prototype the automation and connected habitat arrays that Challenge B requires, and scale them up on early signals of high performance.

### Challenge B: is evolution forecastable in a complex spatial environment?




If the programme progresses to Challenge B, we will run it for at least two six-month rounds, subject to programme continuation at the gates. Populations occupy a grid of connected habitats. Stress is constant within each habitat but rises across the grid, from none to lethal, and worms migrate between neighbouring habitats. As populations adapt, they colonise harsher habitats. At the expanding edge, they shrink and lose genetic diversity, as real populations do at range margins. Teams forecast which habitats are colonised, where expansion stops, and how gene variants distribute across the grid. Challenge B can run as a continuing experiment with successive forecasting rounds. At each sampling point, new data are released and teams forecast both population trajectories and genetic change over the next 22 generations. Each round therefore tests both kinds of forecast over the same period.

We will open negotiations with transfer partners (TA4) during stage 1, and commit to specific Challenge C designs on early signals of high performance in Challenge B.

### Challenge C: is evolution forecastable in real-world contexts?

We will run transfer challenges with partners who hold or can generate genomic time series in applied systems. We have already spoken with potential partners in several industrial contexts.

To protect partners' IP, we could bring algorithms to their datacentres, with models running inside controlled environments and only agreed predictions and evaluation scores released. Raw data would remain with the partner. Historical samples could be sequenced under coded identifiers, with later time points withheld from model training and used for blind evaluation.

## Gates

Two kinds of decisions shape the programme. Gates decide whether it continues. Round results decide what runs next.

Kill gates stop the programme if it is not delivering:

+ Gate 0 (month 6): feasibility. We stop if we cannot rear, stress, count and sequence populations at the required scale, pace and cost, if >15% of populations collapse in the absence of a stressor, or if genetically diverse populations persist no longer than inbred ones under the same pressure.
+ Gates 1 and 2 (months 18 and 30): performance. We stop if the leading forecasts improve on the baseline by less than 1%.

Baseline: initially, the better of a genetics-free demographic model and the best non-genomic quantitative-genetics model, commissioned and maintained through TA2, with suitable competitor-contributed models also eligible. Subsequently, the best model from the previous round. For continuation gates, improvement is the reduction in population size forecast error relative to the baseline, measured using the weighted interval score described in the Annex. Genetic forecasts are assessed separately using the genetic scoring rules. We will refine the kill and advance thresholds using Stage 0 simulations and power analysis, and publish them before Challenge A opens. On entering a new challenge, TA2 will specify a suitable reference model before the round opens.

Advance decisions follow every six-month round:

+ Advance: after at least two completed rounds of the current challenge, the leading population size forecast correctly calls at least 75% of survivals and 75% of collapses among stressed populations (Challenge B: 75% of colonised and 75% of unoccupied habitats), and outperforms the baseline. The next round moves to the more complex challenge, and the previous challenge stops. Challenge B can start from month 18 at the earliest; Challenge C from month 30.

+ Otherwise: another round of the same challenge. Teams learn from the released data, and we refine the design. Only a gate can stop the programme.

## Why a competition, and what it requires

Open competition has repeatedly brought outside methods into biological problems. CASP gave protein structure prediction a common scoreboard; CAFA (Critical Assessment of protein Function Annotation) extended it to protein function; DREAM (Dialogue for Reverse Engineering Assessment and Methods) applied it to gene networks, expression and drug response. In each case competitors predict a fixed feature of a system, rather than forecasting how that system changes. And each was possible only because a large dataset already existed. Evolutionary biology has no equivalent of the Protein Data Bank, on which AlphaFold was trained: a large, standardised, openly accessible record of real outcomes to train and test against. Our programme builds one.

A forecast can only be trusted if three conditions hold.

1. It must be tested against real biological outcomes rather than simulations.
2. The outcomes must be withheld from forecasters until forecasts are submitted.
3. There must be enough independent tests to separate a well-calibrated forecast from a lucky guess.

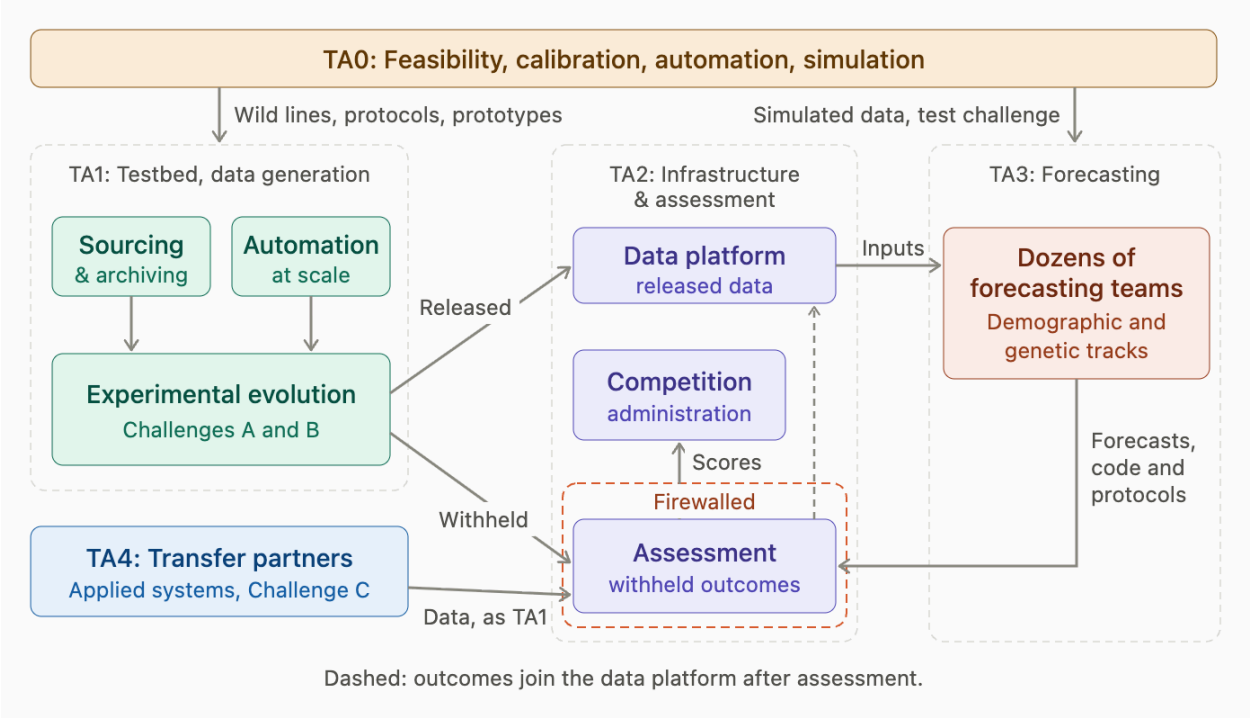
We add a fourth, specific to evolution:

4. forecasts must hold across varied scenarios, to show a model has extracted the rules of the system rather than memorised an analogous case.

## What we expect to fund

Turning evolutionary forecasting into a testable discipline requires:

+ **TA0**: Establishing feasibility and calibrating the experiments.
+ **TA1**: Running evolution experiments at scale and measuring their outcomes.
+ **TA2**: Competition machinery that controls what forecasters see and when, holds the withheld outcomes, and scores submissions against them.
+ **TA3**: Teams blindly forecasting those outcomes, who work from the same well-characterised evidence and testing conditions.
+ **TA4**: Partners to help us create competitions to see how far forecasting transfers to real-world systems




We expect to fund these activities over 42 months, with continuation subject to the gates described above. The competitions are designed to be accessible to teams coming from outside the life sciences.

## What we are still trying to figure out

- Are there technical constraints, failure modes or unintended incentives in the proposed continuation and advancement gates that we have overlooked?
- What amounts of cash should we allocate to small, innovative forecasting teams we want to support?
- Who else might complement our efforts to make this prediction competition successful?
    - Might cloud computing providers/hyperscalers host datasets and/or contribute compute credits? This could help attract users to their platform.
    - Might someone provide sequencing capacity?
    - Might an organisation want to contribute to prize pots or give stipends to creative teams with high potential? Supporting this initiative could complement internal R&D to make the field advance faster.
- Which data formats should datasets be shared in? Current thinking is that we provide some or all of:
    - Reference genome

    - Raw Illumina reads
    - Read alignments
    - Variants / pooled allele frequencies
    - mRNA abundance
    - Single-worm phased genomes
- Teams will be evaluated based on differences between predictions and realised outcomes. The relative rankings of teams will be shared among competition participants. However, to help prevent cheating, teams should also deposit the materials from which the predictions were generated (protocol, model code, links to training data) to TA2's firewalled assessment function under NDA. The verification team should audit or verify competitors' submissions in a manner with no unnecessary disclosure, in order to preserve competitors' IP and commercial confidentiality.
- To focus innovation on modeling approaches, we suspect that a rule similar to "using a widely accessible closed-source dataset or tool" is ok (e.g., broadly available closed-source closed-weight LLMs). But that "using a dataset my company has exclusive access to" should not be.

## SOURCES

*References cited in this document.*


1. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

2. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).

3. Brixi, G. *et al.* Genome modelling and design across all domains of life with Evo 2. *Nature* **652**, 1349–1361 (2026).

4. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).

5. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).

6. Kachroo, A. H. *et al*. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921–925 (2015).

7. Whiting, J. R. *et al*. The genetic architecture of repeated local adaptation to climate in distantly related plants. *Nat. Ecol. Evol.* **8**, 1933–1947 (2024).

8. Cheng, J. *et al*. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

9. Pentz, J. T., Biswas, A., Alsaed, B. & Lind, P. A. Extending evolutionary forecasts across bacterial species. *Proc. Biol. Sci.* **291**, 20242312 (2024).

10. Fusca, D. D., Dall'Acqua, M. N., Sánchez-Ramírez, S. & Cutter, A. D. Phylogenomic timetree-calibrated speciation clocks for Caenorhabditis nematodes reveal slow but disproportionate accumulation of post-zygotic reproductive isolation. *PLoS Genet.* **21**, e1011852 (2025).

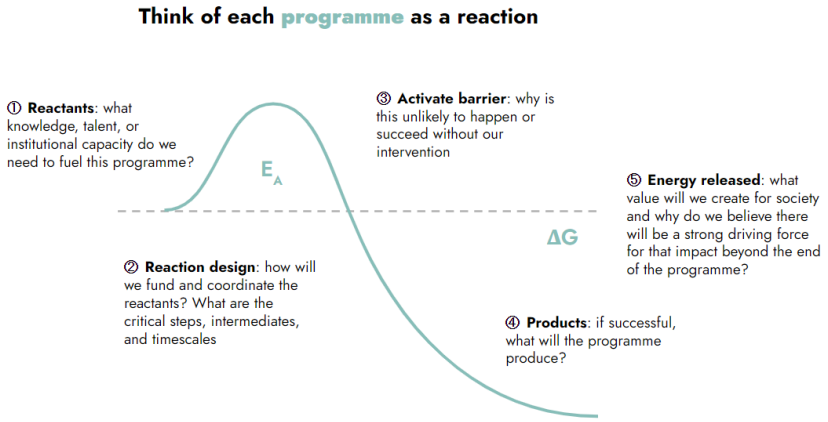
11. Hendry, A. P., Day, T. & Taylor, E. B. Population mixing and the adaptive divergence of quantitative traits in discrete populations: a theoretical framework for empirical tests. *Evolution* **55**, 459–466 (2001).

12. Durkee, L. F., Olazcuaga, L., Melbourne, B. A. & Hufbauer, R. A. Immigration delays but does not prevent adaptation following environmental change: experimental evidence. *J. Evol. Biol.* **37**, 665–676 (2024).

13. Polechová, J. & Barton, N. H. Limits to adaptation along environmental gradients. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6401–6406 (2015).

## PROGRAMME THESIS, REACTION DIAGRAM SUMMARY

*We can metaphorically think of an ARIA programme as a chemical reaction. We present a simple reaction diagram to summarise the key elements of the imagined programme.*




Forecasting Evolution

---

① **Reactants:** The UK holds two halves of this reaction and has never put them together. Frontier AI laboratories, specialist AI-for-biology companies and emerging evolutionary-data platforms build models of sequence, structure, and dynamics, but rarely meet evolution as a specified forecasting task. Furthermore, the performance of their models would likely improve with improved models for evolutionary processes. Alongside them sit commercial breeders, pharmaceutical and agrochemical companies, and statutory agencies making billion-pound decisions whose outcome evolution determines. The UK's communities in quantitative genetics, animal and plant breeding, evolutionary ecology and experimental evolution are long-established and world leading.

② **Reaction Design:** ARIA generates unprecedented-scale evolution experiments, initially in controlled conditions, where populations face increasing heat or chemical stress. This creates a forecasting problem with two measurable answers: which (of thousands) of population survives, and how do the frequencies of its genetic variants shift. Demographic forecasts use earlier census, genome and expression data. Genetic forecasts additionally use released demographic outcomes, while later molecular outcomes remain withheld.

Forecasting scenarios will become successively more complex through the programme: a simple habitat, then a spatial grid of connected habitats, then real-world systems. The competition format supplies innovative modelling approaches. Compute and development awards make it easier for small UK teams with promising ideas to compete.

③ **Activation barrier — market failure**: Forecasting evolution is a public good: its benefits would

be diffuse, long-term and shared across medicine, agriculture, biosecurity, and conservation, while the costs are concentrated upfront. No laboratory, company or conventional funding programme has the remit or incentive to build the experimental scale, automation, repeated molecular measurements, compute, storage and sustained data presentation required. Conventional mechanisms also fail to fund the organisational work: translating the biology for machine-learning teams, recruiting communities that have never worked in this area, and running a credible competition. ARIA can absorb the costs and thus lower a new pathway.

④ **Products:** An unprecedented, machine-learning-ready open dataset of thousands of laboratory evolutionary trajectories, spanning genomes, expression, phenotype and population fate across generations. Forecasting models that outperform non-genomic reference models and transfer to new settings. Forecasting challenges in real-world systems, run with partners on their genomic time series from settings such as pest-resistance monitoring or serial tumour sequencing, which measure how far the models, measurements and rules learnt in worms carry over. A standing benchmark, scoring framework and automated experimental platform; and a community capable of extending the applications to many other life science contexts.

⑤ **Energy released:** Evolution determines the impact of more than £1 trillion of decisions each year. Reliable multigenerational forecasts remain unavailable for many of these decisions. Reliable forecasting would let breeders, clinicians, regulators and agencies act before resistance, collapse or escape, rather than after. Even an imperfect result is worth having too: establishing that the honest forecast horizon is, for example, five generations, would improve decision making throughout those sectors. Either way, the UK holds the dataset, the benchmark and the evidence needed to guide further investment.

# ANNEX

We welcome technical feedback on experimental assumptions, reference models, and the proposed scoring rules provide a rigorous test of forecasting skill.

## Competition challenges and measures of prediction

+ **Forecasting population size:** for each challenge there will be a prize for forecasting population size. Accuracy in forecasting survival and collapse determines progression to the next Challenge; the kill gates use skill relative to the baseline. This skill is also closely aligned to the programme North Star.
+ **Forecasting genetic composition:** the parallel competition for teams forecasting the genetic changes will run 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. Later-generation genomic and gene expression outcomes remain withheld until genetic submissions close.

### Forecasts

**Population size (demographic) forecasts**
For challenges A and B.

#### What is forecast?

For each selected population, teams forecast its size in every generation from the data release to the end of the current forecast window. Size is the number of larvae the population itself produces, counted after selection but before the generation's gene flow, so that the quantity forecast is the population's own productivity rather than productivity plus immigration. A population that produces no larvae in Challenge A is extinct, receives no further migrants and does not recover. Challenge B instead tracks habitat occupancy under migration between neighbouring habitats; an unoccupied habitat may subsequently be colonised.

Two quantities follow from this trajectory, and are scored separately because they answer different questions: how large the population will be, and whether it will still exist.

#### Why a distribution, not a number?

The same argument applies as for genetic forecasts (below). Even an exceptional model will not generate a perfect forecast. Forecasters therefore state their expectation as a distribution, and are scored on its calibration.

Submitting a complete predictive distribution for every generation and thousands of populations would be impractical. Teams instead submit the median and the bounds of nested 80%, 90% and 95% prediction intervals: seven numbers per population per generation.

### Scoring population size

Population sizes span orders of magnitude, from tens of thousands to extinction. An error of 500 individuals has very different meanings at those extremes, so forecasts are made and scored on a log scale, matching the multiplicative nature of population growth. Counts are transformed as log(1 + n), so that extinction is represented as zero on the scored scale rather than being undefined. The same transform is applied to the reference model, so that skill compares like with like.

Scores use the weighted interval score (WIS), a quantile-based approximation to the continuous ranked probability score. It is the continuous counterpart of the ranked probability score used for genetic forecasts, allowing the programme to use one family of scoring rules throughout: Brier score for binary outcomes, RPS for ordered categories, and WIS and CRPS for continuous outcomes.

Each interval score penalises both the width of the interval and how far the observed value falls outside it. WIS combines these scores using specified weights. It also encourages a team to report its honest predictive distribution.

### Scoring extinction

For Challenge A, survival is scored separately with the Brier score, using each team's forecast probability that a population will be extinct by each generation. For Challenge B, the Brier score assesses the forecast probability that a habitat will be unoccupied at each generation. This is the two-outcome special case of the ranked probability score.

Separate extinction scoring matters because the two tasks can come apart. A forecast may track a declining trajectory well but miss the generation in which the population reaches zero; for the applications in this programme, that transition is often the decision-relevant outcome.

In Challenge A, once a population is extinct, it receives no further migrants and remains at zero. Those later generations are therefore excluded from the size score; size is scored up to and including the generation of extinction. In Challenge B, unoccupied habitats remain in the forecast and scoring set because migration can lead to subsequent colonisation.

Reference model and skill

Raw scores are reported as skill relative to a reference forecast. For the first round, the reference is whichever of two models developed without genomic data scores better:

1. A published, deliberately genetics-free reference model: a density-dependent demographic model fitted to the released data from the training generations, that knows the imposed mortality schedule and migration rate but has no genetic data.
2. The best quantitative-genetics model that uses no genomic information, commissioned and maintained through TA2, with suitable competitor-contributed models also eligible.

$$WISS = 1 - \frac{\Sigma_i\, WIS_{(team,i)}}{\Sigma_i\, WIS_{(reference,i)}}$$

The sums are taken before calculating the ratio. For subsequent rounds, the reference is the best model from the previous round, evaluated on the current round's withheld outcomes, as for the programme gates. A skill score of zero means no improvement on that round's reference; one is perfect prediction; and a negative score means that the team performed worse than the reference.

### Genetic forecasts: challenge A:

Across the genome there are millions of positions where individuals in the base population differ. At each such position (locus), the population has a frequency: the proportion of copies that carry one version of the sequence rather than the other. Evolution, reduced to its core, is the change in those frequencies over generations. The forecasting task is therefore to state, for all initially polymorphic loci, what those frequencies will be during the experiment—and, crucially, how confident the team is.

Confidence matters because even a perfect model cannot return a single answer. Which individuals happen to survive and reproduce is partly stochastic, so two identical populations experiencing identical selection will finish with different allelic frequencies across the genome. That spread is a property of the system, not a gap in our knowledge, which is why we ask teams for a range of outcomes with probabilities attached rather than a single prediction. A team that can justify a narrow range is claiming more, and is scored accordingly.

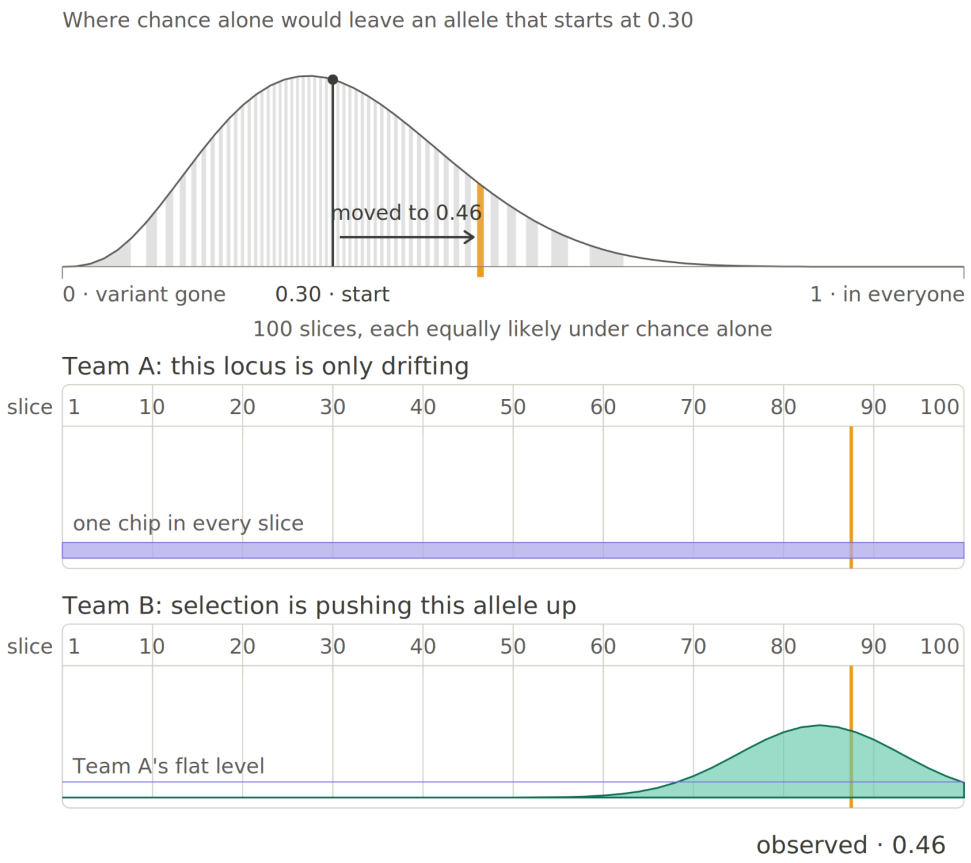
We therefore ask teams to submit a distribution, not a number, using a betting board. For each position we take the frequency the base population started with and simulate what chance alone would do to it—drift plus the gene flow the experiment imposes each generation, with no selection. That simulated spread is divided into 100 slices of equal

probability. Because the neutral outcome is far more likely to be near the starting frequency than far from it, these slices are narrow in the middle and wide at the extremes.

This construction sets the baseline. A team with no information should place one chip in each slice, because under chance alone every slice is equally likely. Any departure from a flat bet is a positive claim that selection is acting—and the further the chips are pushed toward a tail, the stronger the claim.

The following illustration shows two teams at the same position. Team A spreads evenly: its forecast is that nothing but chance is operating here. Team B stakes heavily on the high-frequency slices: its forecast is that the variant is favoured by selection. When the experiment is unblinded, the realised frequency falls in slice 88. Team B, having committed to the right region, scores well. Had the outcome landed near the middle, Team B would have been penalised for their misplaced confidence, while Team A's flat bet would have scored exactly zero—no better and no worse than the null.

Scores are summed across all positions in the scored set. Because the board is built from the neutral model of our experiment, a skill score of zero always means "no better than assuming chance", whatever the position's starting frequency or the population's history.




atever

**Forecasting an allele frequency.** For each variable position we simulate what the experiment would do to it with no selection acting—drift plus the gene flow imposed each generation—and cut the resulting spread into 100 equally likely slices. Because a neutral outcome is far more likely to sit near the starting frequency than far from it, the slices are very narrow there and very wide at the extremes. Teams distribute 100 chips across the slices. One chip in every slice is the null forecast, worth zero; concentrating chips is a claim that selection is acting, and the further into a tail they go, the stronger the claim. Here the allele began at 0.30 and finished at 0.46—a rise chance alone would seldom produce. Team A forecast drift and scores nothing. Team B forecast an upward sweep and is rewarded, having accepted a below-null stake across slices 1 to 65 to buy conviction above them.

### The Scoring System

A team's submission for one position is a set of one hundred numbers, $p_1 \dots p_{100}$, its chips divided by one hundred, summing to one. These are probabilities: p  is the team's expectation that the realised frequency falls in slice j. The flat submission, one chip everywhere, corresponds to p  = 0.01 throughout.

Because the slices are ordered—slice 1 is the lowest frequency, slice 100 the highest—we score against the cumulative forecast rather than the individual slices. Write P  = $p_1$ + … +

p  for the team's cumulative probability up to slice j, and O  for the observation: zero for every slice below the one the outcome fell in, one from that slice onward. Because the final slice in a cumulative probability is always 1, the *ranked probability score (RPS)* for that position is:

$$RPS = \frac{\sum_{j=1}^{j=99} \left(P_j - O_j\right)^2}{99}$$

Lower is better; the score runs from 0 (a forecast that puts everything in the correct slice) to 1 (a forecast that puts everything at the opposite end of the board).

Four properties make this the right rule for this competition:

+ *A forecaster minimises their expected score by reporting exactly what they believe*. A team minimises its RPS value by submitting the distribution it actually believes. Calibration becomes the winning strategy.
+ *It incorporates the magnitude of forecast failure*. A team whose chips sit one slice from the truth has made a nearly correct forecast; a team whose chips sit at the far end of the board has not. A rule that scored only the chips landing on the correct slice would treat these as identical.
+ *It is bounded*. Summed over millions of positions, an unbounded rule allows a single locus where a team assigned near-zero probability to the realised outcome to overwhelm its entire submission. Bounding prevents this.
+ *Connection to Brier score*. When a forecast event has only two possible outcomes, the RPS becomes identical to the standard Brier Score, as will be used for forecasting population persistence.

Reporting

RPS values are not interpretable on their own, so we report a skill score against the flat forecast. Summing over all scored loci, *i*, the *ranked probability skill score (RPSS)* is:

$$RPSS = 1 - \frac{\Sigma_i\left(RPS_{team,\,i}\right)}{\Sigma_i\left(RPS_{reference,\,i}\right)}$$

The sums are taken before the ratio, since averaging ratios locus by locus is biased. A score of zero means the team did no better than assuming chance alone; one is a perfect forecast; a negative score means the team would have done better submitting nothing. Worked example: scoring an evolutionary forecast

The competition uses one hundred slices. This example uses five, so the arithmetic can be followed by hand. Nothing else changes.

## Placing bets

Five slices, each equally likely under chance alone. Four teams each stake 100 chips. The experiment runs, and the realised frequency falls in slice 4.

| Team | 1 | 2 | 3 | 4 | 5 | What they are saying |
|---|---|---|---|---|---|---|
| A | 20 | 20 | 20 | 20 | 20 | Nothing but drift |
| B | 5 | 5 | 10 | 70 | 10 | A sweep, landing around slice 4 |
| C | 5 | 5 | 10 | 10 | 70 | A sweep, slightly stronger than B thinks |
| D | 70 | 10 | 10 | 5 | 5 | The allele is being selected against |

Team B is right. Team C is one slice out. Team D is badly wrong.

### Step 1: Conversion to cumulative forecasts

Divide each team's chips by 100 and accumulate from left to right. Only the first four are needed, because the fifth is always 1.

| Team | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|
| A | 0.20 | 0.40 | 0.60 | 0.80 |
| B | 0.05 | 0.10 | 0.20 | 0.90 |
| C | 0.05 | 0.10 | 0.20 | 0.30 |
| D | 0.70 | 0.80 | 0.90 | 0.95 |

### Step 2: The real outcome, expressed cumulatively

The outcome fell in slice 4, so the cumulative observation is 0 below that slice and 1 from it onward:

$O_1 = 0 \quad O_2 = 0 \quad O_3 = 0 \quad O_4 = 1$

### Step 3: Average the square of the differences

The ranked probability score is the mean squared difference between the cumulative forecast and the cumulative observation. Taking Team A:

$$(0.20 - 0)^2 + (0.40 - 0)^2 + (0.60 - 0)^2 + (0.80 - 1)^2$$

$$= 0.04 + 0.16 + 0.36 + 0.04 = 0.60$$

Divide by 4, giving RPS = 0.150.

| Team | Sum of squares | RPS | Skill score |
|---|---|---|---|
| A | 0.600 | 0.150 | 0.00 |
| B | 0.0625 | 0.0156 | +0.90 |
| C | 0.5425 | 0.1356 | +0.10 |
| D | 1.9425 | 0.4856 | –2.24 |

The skill score is 1 – RPS / RPS(flat). Team A is the flat forecast, so its skill is zero by construction. Team B is close to perfect. Team D has done far worse than saying nothing.

**Evolutionary forecasts: challenge B:**

In Challenge B, populations occupy a grid of connected habitats, so an allele has a frequency in each habitat rather than one frequency overall. Teams forecast that whole spatial distribution: for each locus, a predictive distribution of its frequency in every habitat.

Submitting a hundred-slice board for every locus in every habitat would be impractical, so teams instead submit two parameters per locus-habitat: the mean of their predictive distribution and its standard deviation, or equivalently the two parameters of a Beta distribution.

Scores use the **continuous ranked probability score (CRPS)**, the continuous limit of the ranked probability score used in Challenge A. It is strictly proper, so calibration remains the winning strategy. It is bounded on [0, 1], so no single locus can dominate a submission. And it has no denominator, so habitats in which an allele is lost or fixed — the expected outcome at range margins — are scored as ordinary events rather than producing infinite penalties.

As in Challenge A, we report skill against a reference: the spatial neutral model, which propagates the starting frequencies forward under drift and the imposed migration between habitats, with no selection. A skill score of zero means no better than chance and migration alone.

## Forecasting Challenges

The experimental design requires a tractable laboratory organism with a short generation time. Having two sexes and obligatory sexual reproduction means that selection with gene flow is implemented, mimicking real-world challenges in which severe selection is localised in part of the species range.

Challenges A and B will use new lines specially collected for this programme from wild populations to retain the naturally high genetic diversity and low linkage of this species.

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)subject to two selective agents calibrated to maintain 40% mortality. The large base population ensures that 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—reducing genetic drift

sufficiently to detect the contribution of over 100,000 loci each contributing a small effect to the genetic variance.

**Challenge A**
Will expose the same genetic variants to selective challenges in different combinations. After 10 generations of adaptation to the culture conditions, each of ten lines 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 generation 0 (population founding) to generation 7, selection will be at an intensity which kills 40% of individuals, after which the intensity will increase at the rate determined by the warm up experiments, with the target of 50% replicate extinction before generation 22. Every generation will be archived by freezing for future investigation and as insurance against experimental disasters.

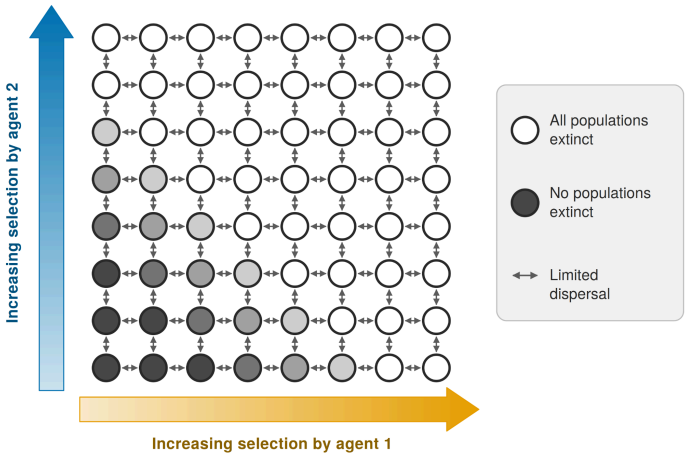
Data generation will comprise

- genomic sequences (9,000x for selected populations, 90,000x for the base);
- expression data for synchronised day-4 adults (polyA+ RNA).

Samples are taken at generations 0, 4, 7, 12, 17 and the final generation, as described above; only data through generation 7 are initially released. Each wild derived base population will be characterised by assembling the genome of 1,000 individuals using Picogram input multimodal sequencing (PiMmS) sequencing offered by the Tree of Life initiative at the Sanger in support of this project.

**Challenge B**
This requires the teams to extend their forecasts to the evolutionary processes that allow species to broaden their environmental niche, and thus spread into novel environments.

The same source populations as used in Challenge A will be introduced across an array of 'habitats', each of which represents a breeding population whose persistence and genetic evolution will be the subject of the forecasting



challenge, as before. Importantly, these habitats are not isolated: each generation there will be limited demographic exchange between adjacent habitats. Timing and duration will depend on programme progression. Two sampling events, 22 generations apart, can be taken in a six month period.

Artificial selection will be fixed for each habitat, with selection intensities of both agents spanning a range from no artificial selection to lethal conditions. Over time, viable (i.e., sustained) breeding populations are expected to establish in the more extreme environments. At the range margins we expect to see loss of genetic diversity[113], 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. The experiment may run for up to 88 generations, subject to continuation and advancement decisions. Forecasting teams will predict which populations produce offspring that survive the selective challenge (a list of populations for each generation in a 22-generation time course), and the distribution of each allele among the populations. Competitions will be run at 22 generation intervals after the release of genomic, transcriptomic, and census data at generations 22, 44 and 66.

## Q&As

### Why selection with gene flow?

The gene flow from very large base population 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 mating events 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 restrains 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 20 replicates of each line and selection regime?

Our target is to achieve sufficient precision to capture the majority of the genetic variance: that requires detecting small effects of common alleles, and larger effects for rarer alleles. With $\beta$ representing the effect size expressed as the proportion of the SD of genetic variance, we aim to detect at 80% power $\beta$=0.005 for a variant locally common (20%) in one base population. Such a small but common effect would explain a tiny fraction of the genetic variance: if spread evenly it would require 125,000 such loci to explain the total

genetic variance. Across all base populations we would aim to detect $\beta$=0.01 for a rare allele (1%), which would contribute an even smaller proportion (requiring 505,000 loci). Using conventional automation and rearing (9cm Petri dishes) we can achieve two selection regimes each 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). Under sustained migration the frequency of each genetic variant rapidly approaches an equilibrium. Under an additive model, allele frequencies will be clustered around a mean displaced from the base-population frequency. The size of this displacement depends on the strength of selection and migration, while the spread around the mean depends primarily on random genetic drift, determined by effective population size, and migration. The replication required was calculated assuming additive genetic effects, the specified allele frequency and effect sizes, a heritability of 0.6, 40% selective mortality, an effective population size of 60,000 and 10% migration per generation, with sequencing depths of 9,000× per selected population and 90,000× for the base population. This calculation assumes constant mortality and approximate migration-selection equilibrium. Stage 0 simulations will assess power during stress escalation and population collapse.