ORIGINAL ARTICLE

Journal Section

Computational resources for simulating under a spatial coalescent model across heterogeneous landscapes and testing hypotheses about the geography of genetic variation: QUETZAL-EGGS, -CRUMBS, -NEST and DECRYPT

Arnaud Becheler[1*] | **L. Lacey Knowles**[1†]

[1]Ecology and Evolutionary Biology Department, University of Michigan, 48109, MI, U.S.

Correspondence
Lacey L. Knowles, Ecology and Evolutionary Biology Department, University of Michigan, 48109, MI, U.S.
Email: knowlesl@umich.edu

Funding information
This study was funded by NSF [DEB 16-55607 to Lacey L. Knowles

Spatially explicit coalescent models in which the underlying demographic parameters are informed by the environment (either past, present, or temporally and spatially chaning environments) provide a framework for hypothesis testing that incorporates geographic information about genetically sampled individuals. This general approach - Integrated Distributional, Demographic and Coalescent (iDDC) modelling - can be used to explain how heterogeneous, dynamic landscapes shape the history and genetic patterns of a species. However, iDDC approaches involve long and complex tasks that often require custom-fit simulators, some coding expertise, and extensive computing resources. Here we introduce several resources that offer improved speed and generality, as well as expand the feasible parameter space for conducting iDDC analyses compared to other software applications. Specifically, QUETZAL-EGGS are C++ iDDC simulators; QUETZAL-CRUMBS is a complementary set of

[*] A.B. designed the models, software and computational framework.
[†] L.L.K. supervised the conception, funding and findings of this work.
[*†] All authors provided critical feedback and helped shape the research, resources and analysis. All authors contributed to the final manuscript.

Python tools for simulating on specific landscapes and conducting Approximate Bayesian Computation (ABC) analyses (e.g., prior sampling, geospatial operations, ENM/SDM, visualization); DECRYPT is a framework for automated, biology informed robustness analysis of the multispecies coalescent model. All these tools and their dependencies for local use or remote computations are made readily available in a Docker container package called QUETZAL-NEST.

KEYWORDS
environmental niche modeling, coalescence, biogeography, software, simulation, landscape

1 | INTRODUCTION

Integrating distributional, demographic and coalescence models (*iDDC modeling*, He et al., 2013, see also the excellent review by Larsson et al. 2021) is a powerful tool to explore how spatial and temporal landscape heterogeneity shapes the genetic diversity of modern populations (*e.g.*, Knowles and Alvarado-Serrano, 2010; Brown and Knowles, 2012; Pan et al., 2020). In this category of modelling approaches, the landscape is discretized into a very large number of demes (e.g., >1000). The demographic history (i.e., the number of individuals in each deme and the number of migrants across demes) is simulated as a function of the environmental variation over many generations (i.e., thousands, to tens of thousands, of generations). Then conditionally on this historical demographic processes, a coalescence process tracks the ancestry backward in time of genetically sampled individuals across a landscape.

When coupled with simulation-based inference methods like Approximate Bayesian Computation, ABC (Beaumont et al., 2002; Estoup et al., 2010), these iDDC models have the interesting property to generate complex geographic distributions of genetic variation while maintaining a reduced number of parameters (because parameters describe landscape-wide processes as a function of the underlying environment, rather than excessive parameterization of each individual deme). With less than a dozen parameters, the procedure is nevertheless flexible enough to represent reasonably complex processes (e.g., shifting species distributions, varied rates of migration across a landscape, population growth, and geographic barriers that vary in their attenuation of gene flow).

However, overall contributions of iDDC modeling have been rather limited. Few researchers apply the approach despite its intriguing potential for hypothesis testing using biologically informed expectations, and even though iDDC modeling addresses questions that could not otherwise be addressed with generic models that are not spatially explicit (e.g., the contribution of contemporary versus historical landscapes to genetic structure; He et al., 2013) recolonization of river routes following deglaciation (Neuenschwander et al., 2008); the geographic position of refugial populations (Bemmels et al., 2019); the facilitative versus competitive effects of co-distributed species on colonization of landscapes (Ortego and Knowles, 2020). We argue that the limited traction of iDDC modeling reflects technical and practical challenges of iDDC modeling itself. To increase the accessibility of iDDC modeling to a broad audience, we have developed a set of software tools that solve some of the methodological hurdles associated with ease of application and computational constraints. Rather than presenting a biological application of these resources in this article, we instead direct readers to an example repository with a full data analysis that is accompanied by detailed documentation

of the applied software and computational resources `https://github.com/Becheler/quetzal_on_OSG`.

2 | QUETZAL-EGGS SIMULATORS

2.1 | Motivations

With respect to available tools for simulating spatially explicit genetic variation across a landscape, SPLATCHE (Currat et al., 2004, 2019) is a user-friendly simulation software that has been supporting the community for two decades, but it is closed source and limited in its configuration capabilities. Considering the wide range of systems that can potentially be analyzed using iDDC, there is not a one-size-fits-all solution: different systems will inevitably require different sets of assumptions/models/simulators. This is exemplified by the many modified versions of the program SPLATCHE used across the literature (*e.g.*, White et al., 2013; Mona et al., 2014). Moreover, because the code is closed source, modifications are restricted to a limited number of people who work with the program and their ability (and availability) to incorporate new implementations.

To encourage the open-source creation of new simulators and foster the analysis of new biological systems, QUETZAL-COATL (Becheler et al., 2019; Becheler and Knowles, 2020) was designed as a C++ library of generic components that can be programmed and assembled into versatile simulators. However, its use is by definition restricted to C++ programmers (although online tutorials may shorten the beginners learning curve). To widen the range of models available to non-programmers, and incorporate information about the landscape for informing the spatial coalescent, we introduce the open-source QUETZAL-EGGS (https://github.com/Becheler/QUETZAL-EGGS).

QUETZAL-EGGS contains ready-to-use simulators for implementing different variants of iDDC models. For example, EGG1 has been developed to simulate fine-grain spatial structure in a system of continental islands formed by progressive submersion of the continental shelf as a response to sea level change after the LGM, but whose populations remain connected to the mainland by transient trans-oceanic dispersal (e.g., rafting events), whereas EGG2 has been developed to illustrate climate-driven pulses in matrix connectivity among relatively isolated populations, such as among montane sky-islands systems (e.g., climate-induced elevational distribution shifts). QUETZAL-EGGS programs take as general inputs a configuration file, a geospatial file describing the landscape of interest and its dynamics (generally a suitability raster from an ENM step, or multiple rasters for ENMs from different time periods), and a table of sampling locations (latitudinal and longitudinal coordinates). QUETZAL-EGGS complements other spatial simulation resources (*e.g.*, slendr Petr et al., 2022) by offering a compromise between model complexity and computational efficiency. For example, slendr and its backend SLIM (Haller and Messer, 2019) have features to represent spatial interactions between individuals, but the demographic events have to be compiled into a R object, which is expected to be computationally challenging when countless migrations events happen across a complex landscape during a long period of time, compared with SPLATCHE3 and QUETZAL-EGGS simulators that are compiled in C++, and as such, extends the model/parameters space for spatial simulations. We again note that anyone is welcome to contribute to discussions, or can update and grow this list of historical scenarios by adding new models using the Github *Issues* or *Pull Request* systems, or by contacting the authors.

2.2 | Memory management

One of the significant improvements with our QUEZTAL pipeline regards the computational expense of iDDC modeling. For example, SPLATCHE (Currat et al., 2019) keeps the demographic history on RAM, and as such, individual simulations run faster. However, this comes at a cost of constraining the historical duration and landscape resolution

(i.e., number of demes) to the system RAM capacity. Because RAM is a more limited resource than disk space, this constrains the number of nodes one can request on computing grids, slowing down the whole workflow and leading to very long run times. In response, researchers try to bypass this problem by re-scaling generation time and/or using coarser landscapes (*i.e.* to reduce the number of generations and/or number of demes in the spatial simulation, respectively, see He et al., 2013), but this makes other parameters of the model difficult to interpret (Massatti and Knowles, 2016) and prevents the emergence of a fine-grain genetic structure that is often a desirable property for hypothesis testing. To mitigate computational constraints, QUETZAL-EGGS offers a compile-time option that implements sliding windows that keeps only two active layers (i.e., two generations of the spatially explicit demographic history informed by environmental heterogeneity) on RAM at a time, storing unused layers on disk. This allows longer histories at higher spatial resolutions to be modelled.

3 | QUETZAL-CRUMBS: PYTHON COMPONENTS SUPPORTING QUETZAL-EGGS

3.1 | Motivations

A number of iDDC related procedures are not *per se* the responsibility of the simulation program and would require some coding expertise to implement. Below we describe a new python3 library, QUETZAL-CRUMBS, that gathers procedures of general interest for iDDC modeling using QUETZAL-EGGS, improving the accessibility of iDDC modelling to a broad user base.

3.2 | Visualization of dynamics landscapes

An important part of model choice and calibration is to visually investigate the landscape historical dynamics, whether it is how the candidate model and its parameters affect the demographic history, or how the suitability landscape changes through time. These 2D quantities are represented at each time step by a geospatial regular grid associated to a Coordinate Reference System (a raster). The temporal heterogeneity is represented by stacking these rasters (a multiband raster), where each layer (or band) represents a landscape at a given time period. To visualize how these stacks change through time, the `crumbs.animate` function converts these stacks into GIF or MP4 animations.

3.3 | Preparing the landscape and adjusting the spatial grid properties

In spatial dynamic models, resolution of the landscape is an issue (see *e.g.*, Bocedi et al., 2012): if the resolution is too low (i.e., large environmentally heterogeneous geographic areas represented as a single deme), biological processes may be misrepresented and biases may result. If the landscape resolution is too high, computational costs may make ABC methodology impossible. Likewise, orientation of the spatial grid is a necessary model parameter, but with multiple orientations possible, this decision is made arbitrarily. To deal with these uncertainties, a common practice is to arbitrarily set a North-up orientation for the spatial grid, and manually guess and adjust the landscape grid resolution to fit computational capacities. However, the impact on inference should be carefully assessed and one way to do so is to include the spatial resolution and grid orientation as parameters to be estimated (*e.g.*, Baird and Santos, 2010; Estoup et al., 2010).

The `crumbs.rotate_and_rescale` function allows the rejection of a sample rotation angle/resolution that can not account for the genetic structure of an empirical data set or the simulation walltime is reached. That is, with this

QUEZTAL-CRUMBS function, the user can avoid too coarse or too fine of spatial landscape grids and identify the rotation angle that provides the best fit to the observed geographic distribution of genetic variation.

3.4 | Beyond the squared spatial grid

There are many ways to discretize (tesselate) a landscape. There has been a focus on discrete grids for iDDC modeling partly because SPLATCHE relies on ASCII raster format. However, it is expected that different tesselation models could affect the inference (Baird and Santos, 2010), and consequently, they should be tested. Moreover, considering different tesselations would allow an efficient integration of key data and processes that operate at different scales, such as capturing local micro-refugia without paying the cost of a landscape-wide high resolution (see *e.g.*, Larsson et al., 2021; Randin et al., 2009; Trivedi et al., 2008). Since QUETZAL-COATL embeds abstract libraries like GDAL, the module does not make strong assumptions about tesselation models, requiring only a concept of coordinates, vicinity and distance for sampled individuals/populations. Consequently, different functions to discretize space (like Voronoi tesselations) can be investigated using QUETZAL-CRUMBS; the shapefiles would then be passed on to the QUETZAL-EGGS simulator.

Rectangular landscapes can have counter-intuitive orientations that are not very convenient to work with, when compared to disk (circular) landscapes. To facilitate landscape manipulation and analysis, we implement a function `circle_mask` that fits and cuts a circle with maximal radius around the landscape center coordinate when rotating and re-scaling landscapes.

3.5 | Representation of temporal heterogeneity at fine scales

Despite appreciable progress in accounting for spatial heterogeneity, iDDC studies have focused on a limited number of bands (that is, raster layers) to represent temporal variability (*e.g.*, 1 for static ENM, 2 or 3 for dynamic ENM, see He et al., 2013). This in large part reflects limitations with the available tools for spatially explicit modeling across temporally varying landscapes (Larsson et al., 2021), without some scripting required (*e.g.*, Brown and Knowles, 2012).

To ease this step, the `crumbs.interpolate` function takes a n-bands geoTiff and assigns its first band to generation 0 and its last n band to the simulation maximal generation parameter g (that is, the present). The $n-2$ remaining bands are then assigned to generations in a regular sequence $[0...g]$, or to a specific sequence provided by the user. Using dask (Rocklin, 2015) for parallel computing and larger-than-memory data management, the whole spatial dynamics is reconstructed by interpolating the missing bands (i.e., bands without independent paleoclimatic data; (see Brown and Knowles, 2012), and this temporal heterogeneity can be animated using `crumbs.animate` and passed to a QUETZAL-EGGS simulator for simulating g generations of a spatial dynamic across the landscape. Note that reconstructing a suitability band for every generation may not scale well to the case of long histories in large landscapes. In these cases, the GDAL Virtual Format (VRT driver, `.vrt`) can be used to build a virtual dataset composed from other GDAL datasets with re-positioning; this allows for very large datasets where most of the bands are actually repeated and reused, rather than physically represented in memory.

Rather than interpolating temporal heterogeneity from a few reference paleoclimatic ENMs for iDDC modeling (e.g., Knowles and Massitti 2018), the CHELSA-Trace21k database (Karger et al., 2016) offers high resolution spatio-temporal reconstructions for bioclimatic and elevational data for every century from the present to the LGM (that is, 220 time steps, with a band each 100 years). Using the `crumbs.get_chelsa` function in QUEZTAL-CRUMBS, the database variables are downloaded with a procedure that clips and assembles the the 220 layers into a GeoTiff dataset for the spatial extent of the sampled data points (the user can specify a margin to extend the landscape to the desired

size). This automation reduces memory usage and the resultant GeoTiff datasets can be processed by other QUETZAL-CRUMBS modules and by the QUETZAL-EGGS simulators. Note that the long download step can easily be distributed on cluster grids.

Together these advances provide a seamless, flexible iDDC workflow that is also open to extensions. Specifically ease of the iDDC workflow is made possible by (i) databases with major past climatic reconstructions (*e.g.*, Worldclim, Fick and Hijmans, 2017; PaleoClim, Brown et al., 2018; CHELSA, Karger et al., 2016), (ii) ENM software tools (*e.g.*, SDMToolbox Brown, 2014; the R `dismo` package, Hijmans et al., 2017), and (iii) the QUETZAL iDDC modeling framework (Becheler et al., 2019) that generate and/or accept user provided GeoTIFF files.

3.6 | Automated High resolution SDM reconstruction

Using a shapefile of sampling coordinates, the `crumbs.sdm` module fetches CHELSA-Trace21k layers, crops them to the area of interest, and performs a species distribution reconstruction by automatically fetching presence points from the GBIF database (or user input files of occurrences) using 4 machine-learning classifiers (namely, Random Forest, Extra Trees, XGB and LGBM classifiers) to perform model fitting with a k-fold cross validation for computing accuracy scores. The models are then projected to past CHELSA-Trace21k layers and a geotiff is assembled.

We are aware of the numerous challenges that SDM involves and debate regarding the best way to generate SDMs. Here we traded heavily customized approaches for a more general and reproducible workflow. This enables non-programmers to produce a suitability layer for every century during the last 21,000 years, and supply these 220 layers of spatial dynamics to the QUETZAL-EGGS genetic simulation programs. Despite errors with the suitability predictions that may result from this more general and simplified automation of modeling landscape suitability, the inferred suitability predictions can be transformed by an arbitrary function (whose parameters can be estimated by ABC) to improve the generative fit of a model to observed genetic variation. This approach is adopted here because for time periods in the more distant past, assumptions for generating highly precise and accurate projections may not hold (e.g., niche conservatism and or similar community composition such that the species interactions are stable and therefore the relationship between specific environmental variables and a species distributions does not change over time). This contrasts with practices for short-term (the present or decadal) predictions where a highly-precise model may be desirable.

3.7 | Genetic simulation and conversion tools

Because in their current version QUETZAL-EGGS simulate coalescent trees in a Newick format that is stored in a SQLITE database along simulation parameters, QUETZAL-CRUMBS handles access to the simulation SQLITE database, and includes simulation of independent DNA sequences (using Pyvolve; Spielman and Wilke, 2015), data format conversion, and summary statistics computation (using Arlsumstat; Excoffier and Lischer, 2010).

Note that for parameter estimation, QUETZAL-CRUMBS implements procedures already covered by pre-existing libraries (Wegmann et al., 2010; Mertens et al., 2018) to simplify bash scripting and dependency management for genetic simulation under specified priors.

In addition, for spatially explicit simulations, initialization of the simulations has a geographic component. Sometimes the geographic origin might be specified (*e.g.*, based on the putative location of glacial refugia; see Bemmels et al., 2019). However, in other situations the origin is unknown and has to be inferred (He et al., 2017); the `crumbs.sample` function randomly samples candidate origin coordinates among the terrestrial cells of a landscape file in geoTIFF format.

3.8 | Sensitivity of inference to sampling of individuals

Practical constraints may affect the sampling of individuals across a landscape (e.g., costs of genotyping many individuals or difficulties with being able to collect specimens). However, the sampling scheme itself may impact inferences made from genetic data (Mason et al., 2020). For example, limited sampling of geographically widespread taxa may generate genetic patterns that deviate from coalescent expectations for a single population, and as a consequence, the data might fit better a "multispecies" coalescent, MSC, model (i.e., more than one population lineage). In such cases, the sampling (rather than limited gene flow) would drive support for multiple population lineages, which in turn, is commonly interpreted as support for multiple cryptic species in the parlance of species delimitation (Barley et al., 2018; Sukumaran and Knowles, 2017). Yet, tests for such biased inferences arising from the sampling design are not common.

The program DECRYPT, which uses QUEZTAL-EGGs to simulate a spatial coalescent informed by the environment (i.e., habitat suitability using QUEZTAL-CRUMBS), can be used to test for sensitivities due to sampling. Specifically, simulated data sets from the posterior of a full iDDC model (i.e., pseudo-observed data sets, PODs) are used to assess the robustness of the MSC to possible violation of its assumptions (e.g., restrictions in gene flow arising from environmental heterogeneity). That is, for a particular geographic sample design (the actual geographic coordinates of empirical samples) the inferred number of lineages can be estimated under the MSC. This provides a test of the robustness of the MSC to violations of the models assumptions, such as genetic structure within a species as an artifact of the sampling scheme or due to reduced gene flow because of landscape features.

4 | QUETZAL-NEST

For a non-programmer and newcomer to iDDC modeling, one the first barrier encountered is the diversity and dispersion of tools and methods: identifying, installing, configuring, calibrating and running the required tools is far from trivial, even for simple tests on a local computer, and not to mention runs conducted on a cluster for scalable, reproducible science. A streamlined software solution that alleviates at least some of the complexity of analyses based on a spatial coalescent model is key to broadening the scope of potential users.

Ideally, we will see the emergence of a framework for reproducible iDDC where the practitioner would only have to (i) connect to an HTC grid, (ii) download content from a standardized Github repository of gathered tools and methods for analysis, (iii) upload their own input files, (iv) select and run routine analysis workflows, and (v) retrieve and interpret outputs. Recent advances make some of this path a bit easier. First, the recent developments of ABC-Random Forest (Raynal et al., 2019) now allow scientists to perform ABC inference, bypassing complex and time-consuming aspects of the inference, which enables the design of more standard ABC workflows. Second, the emergence of containers (*e.g.*, Docker and Singularity Kurtzer et al., 2017) and distributed High Throughput Computing (*e.g.*, the Open Science Grid, Pordes et al., 2007) now allow packages to be shared and run in reproducible analytical environments.

As a first step in this direction, we developed the QUETZAL-NEST Docker container that comes with about 65 preinstalled dependencies. The container is published on DockerHub and available for local use (*e.g.*, development, tests, tutorials) with `docker pull arnaudbecheler/quetzal-nest`. To allow researchers to perform full iDDC inferences with ABC, QUETZAL-NEST has also been submitted to the Open Science Grid CVMFS image repository where it is available for distributed High Throughput Computing. An example repository of a full data analysis workflow built for OSG can be found at `https://github.com/Becheler/quetzal_on_OSG`.

5 | FUTURE PROSPECTS

Going forward, landscape types, demographic and historical details, and geographic settings will expand beyond the current resources of QUEZTAL, and will be made available as additional QUETZAL-EGGS simulators beyond the current list. Such additions are eased by the clear structure and intent of the C++ files that define each existing QUETZAL-EGGS: `EGG_options.h` defines the simulator options, `EGG_context.h` defines the forward/backward model, `EGG_database` is responsible for storing simulated parameter values and data, and the `main.cppp` contains the main function. All files are relatively short and the code can be modified with minimal C++ knowledge.

Currently, Quetzal simulates independent loci. Although this assumption is simple, it still matches a large number of existing geospatial genetic datasets. However, it also ignores the rich information embedded in recombination patterns of more complex datasets. Given its open source code and abstract interfaces, Quetzal could be interfaced with TSKIT (Kelleher et al., 2018) and/or SLIM for computationally efficient generated spatial history of whole genomes. More specifically, we began to implement a C++ version of the Hudson algorithm for enabling the simulation of a structured coalescent with recombination.

With this flexibility in mind, we have developed QUEZTAL so it can continue to evolve to fit future demands of spatially explicit genetic studies in an open environment that is available to all researchers.

acknowledgements

We thank two anonymous reviewers for useful comments of an earlier version of the paper. This work was supported by NSF DEB 1655607 (to LLK).

conflict of interest

None declared.

references

Baird, S. J. and Santos, F. (2010) Monte carlo integration over stepping stone models for spatial genetic inference using approximate bayesian computation. *Molecular ecology resources*, **10**, 873–885.

Barley, A. J., Brown, J. M. and Thomson, R. C. (2018) Impact of model violations on the inference of species boundaries under the multispecies coalescent. *Systematic Biology*, **67**, 269–284.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Becheler, A., Coron, C. and Dupas, S. (2019) The quetzal coalescence template library: A c++ programmers resource for integrating distributional, demographic and coalescent models. *Molecular ecology resources*, **19**, 788–793.

Becheler, A. and Knowles, L. L. (2020) Occupancy spectrum distribution: application for coalescence simulation with generic mergers. *Bioinformatics*, **36**, 3279–3280.

Bemmels, J. B., Knowles, L. L. and Dick, C. W. (2019) Genomic evidence of survival near ice sheet margins for some, but not all, north american trees. *Proceedings of the National Academy of Sciences*, **116**, 8431–8436.

Bocedi, G., Pe'er, G., Heikkinen, R. K., Matsinos, Y. and Travis, J. M. (2012) Projecting species' range expansion dynamics: sources of systematic biases when scaling up patterns and processes. *Methods in Ecology and Evolution*, **3**, 1008–1018.

Brown, J. L. (2014) Sdm toolbox: a python-based gis toolkit for landscape genetic, biogeographic and species distribution model analyses. *Methods in Ecology and Evolution*, **5**, 694–700.

Brown, J. L., Hill, D. J., Dolan, A. M., Carnaval, A. C. and Haywood, A. M. (2018) Paleoclim, high spatial resolution paleoclimate surfaces for global land areas. *Scientific data*, **5**, 1–9.

Brown, J. L. and Knowles, L. L. (2012) Spatially explicit models of dynamic histories: examination of the genetic consequences of pleistocene glaciation and recent climate change on the american pika. *Molecular Ecology*, **21**, 3757–3775.

Currat, M., Arenas, M., Quilodràn, C. S., Excoffier, L. and Ray, N. (2019) Splatche3: simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal. *Bioinformatics*, **35**, 4480–4483.

Currat, M., Ray, N. and Excoffier, L. (2004) Splatche: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.

Estoup, A., Baird, S. J., Ray, N., Currat, M., CORNUET, J.-M., Santos, F., Beaumont, M. A. and Excoffier, L. (2010) Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad bufo marinus. *Molecular ecology resources*, **10**, 886–901.

Excoffier, L. and Lischer, H. E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under linux and windows. *Molecular ecology resources*, **10**, 564–567.

Fick, S. E. and Hijmans, R. J. (2017) Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, **37**, 4302–4315.

Haller, B. C. and Messer, P. W. (2019) Slim 3: forward genetic simulations beyond the wright–fisher model. *Molecular biology and evolution*, **36**, 632–637.

He, Q., Edwards, D. L. and Knowles, L. L. (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution*, **67**, 3386–3402.

He, Q., Prado, J. R. and Knowles, L. L. (2017) Inferring the geographic origin of a range expansion: Latitudinal and longitudinal coordinates inferred from genomic data in an abc framework with the program x-origin. *Molecular Ecology*, **26**, 6908–6920.

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J. and Hijmans, M. R. J. (2017) Package 'dismo'. *Circles*, **9**, 1–68.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P. and Kessler, M. (2016) Chelsa climatologies at high resolution for the earth's land surface areas (version 1.0).

Kelleher, J., Thornton, K. R., Ashander, J. and Ralph, P. L. (2018) Efficient pedigree recording for fast population genetics simulation. *PLoS computational biology*, **14**, e1006581.

Knowles, L. L. and Alvarado-Serrano, D. (2010) Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic and genetic models in montane grasshoppers. *Molecular Ecology*, **19**, 3727–3745.

Kurtzer, G. M., Sochat, V. and Bauer, M. W. (2017) Singularity: Scientific containers for mobility of compute. *PloS one*, **12**, e0177459.

Larsson, D. J., Pan, D. and Schneeweiss, G. M. (2021) Addressing alpine plant phylogeography using integrative distributional, demographic and coalescent modeling. *Alpine Botany*, 1–15.

Mason, N. A., Fletcher, N. K., Gill, B. A., Funk, W. C. and Zamudio, K. R. (2020) Coalescent-based species delimitation is sensitive to geographic sampling and isolation by distance. *Systematics and Biodiversity*, **18**, 269–280.

Massatti, R. and Knowles, L. L. (2016) Contrasting support for alternative models of genomic variation based on microhabitat preference: Species-specific effects of climate change in alpine sedges. *Molecular Ecology*, **25**, 3974–3986.

Mertens, U. K., Voss, A. and Radev, S. (2018) Abrox—a user-friendly python module for approximate bayesian computation with a focus on model comparison. *PloS one*, **13**, e0193981.

Mona, S., Ray, N., Arenas, M. and Excoffier, L. (2014) Genetic consequences of habitat fragmentation during a range expansion. *Heredity*, **112**, 291–299.

Neuenschwander, S., Largiadèr, C. R., Ray, N., Currat, M., Vonlanthen, P. and Excoffier, L. (2008) Colonization history of the swiss rhine basin by the bullhead (cottus gobio): inference under a bayesian spatially explicit framework. *Molecular Ecology*, **17**, 757–772.

Ortego, J. and Knowles, L. L. (2020) Incorporating interspecific interactions into phylogeographic models: A case study with californian oaks. *Molecular Ecology*, **29**, 4510–4524.

Pan, D., Hülber, K., Willner, W. and Schneeweiss, G. M. (2020) An explicit test of pleistocene survival in peripheral versus nunatak refugia in two high mountain plant species. *Molecular ecology*, **29**, 172–183.

Petr, M., Haller, B. C., Ralph, P. L. and Racimo, F. (2022) slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes. *bioRxiv*. URL: `https://www.biorxiv.org/content/early/2022/03/21/2022.03.20.485041`. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2022/03/21/2022.03.20.485041.full.pdf.

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., Avery, P., Blackburn, K., Wenaus, T., Würthwein, F. et al. (2007) The open science grid. In *Journal of Physics: Conference Series*, vol. 78, 012057. IOP Publishing.

Randin, C. F., Engler, R., Normand, S., Zappa, M., Zimmermann, N. E., Pearman, P. B., Vittoz, P., Thuiller, W. and Guisan, A. (2009) Climate change and plant distribution: local models predict high-elevation persistence. *Global Change Biology*, **15**, 1557–1569.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P. and Estoup, A. (2019) Abc random forests for bayesian parameter inference. *Bioinformatics*, **35**, 1720–1728.

Rocklin, M. (2015) Dask: Parallel computation with blocked algorithms and task scheduling. In *Proceedings of the 14th python in science conference*, vol. 130, 136. Citeseer.

Spielman, S. J. and Wilke, C. O. (2015) Pyvolve: a flexible python module for simulating sequences along phylogenies. *PloS one*, **10**, e0139047.

Sukumaran, J. and Knowles, L. L. (2017) Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*, **114**, 1607–1612.

Trivedi, M. R., Berry, P. M., Morecroft, M. D. and Dawson, T. P. (2008) Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. *Global change biology*, **14**, 1089–1103.

Wegmann, D., Leuenberger, C., Neuenschwander, S. and Excoffier, L. (2010) Abctoolbox: a versatile toolkit for approximate bayesian computations. *BMC bioinformatics*, **11**, 1–7.

White, T. A., Perkins, S. E., Heckel, G. and Searle, J. B. (2013) Adaptive evolution during an ongoing range expansion: the invasive bank vole (m yodes glareolus) in i reland. *Molecular Ecology*, **22**, 2971–2985.

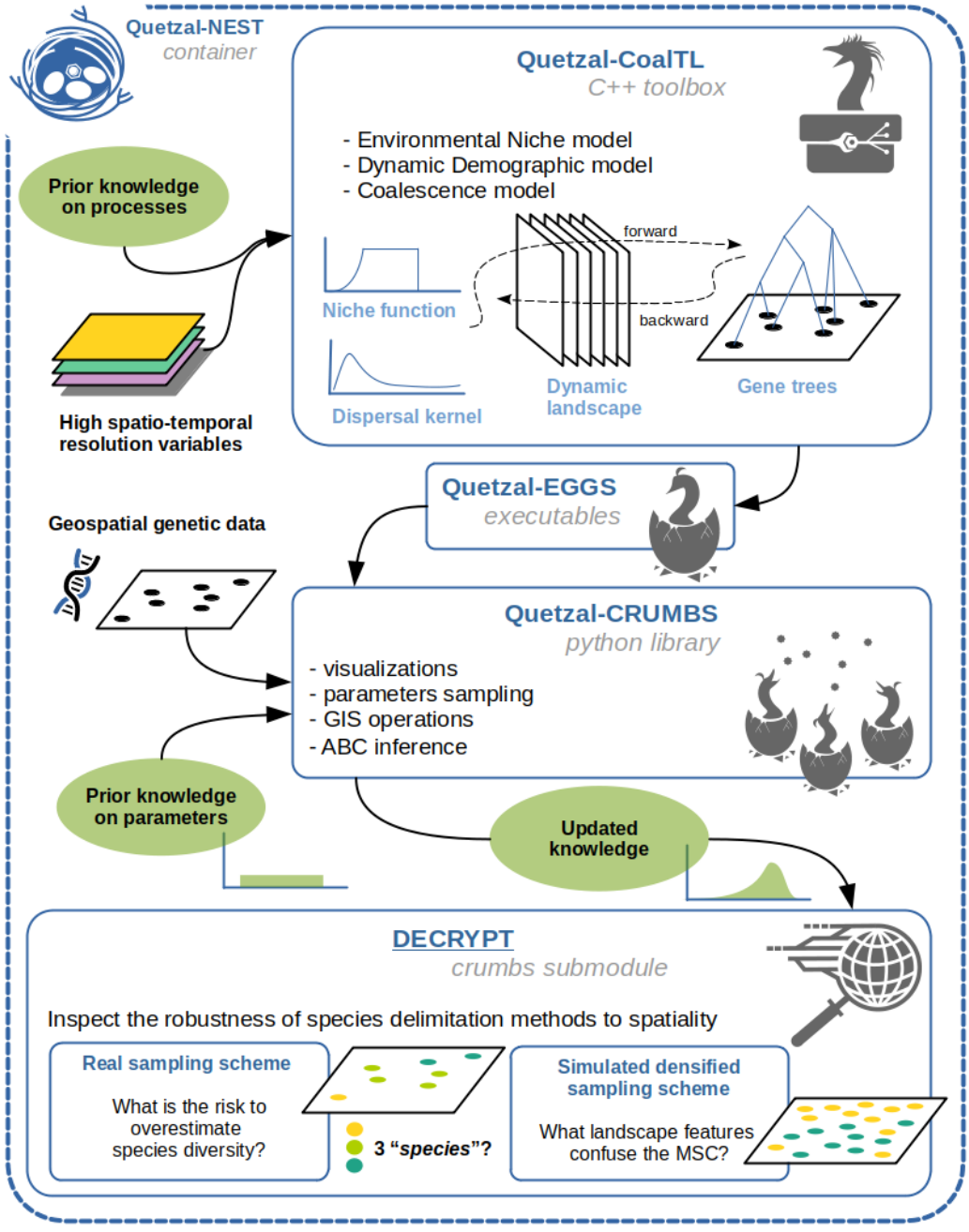

FIGURE 1 Main components, concepts and uses of the QUETZAL framework for open source iDDC modeling. QUETZAL-NEST is a Docker container that packages all the tools and dependencies; it can be run locally with Docker or on dHTC clusters with Singularity. QUETZAL-COATL (Becheler et al., 2019; Becheler and Knowles, 2020) is a C++ library of reusable components and QUETZAL-EGGS are C++ iDDC simulators built with these components. QUETZAL-CRUMBS is a complementary set of Python tools for hypothesis testing using ABC and common landscape-ABC problems, including automatic adjustment of the spatial resolution and orientation of the landscape. DECRYPT is a submodule of CRUMBS for automated, biology-informed robustness analysis of the multispecies coalescent model.