Adding species to stdpopsim

# Outline

## Introduction

1. Simulation is important for population genomics, both methods development and inference
   * there is a building avalanche of genomic data for a variety of species, and a complementary flourishing of methods for population genomic inference
     + so the number and variety of species requiring detailed and accurate simulation models is increasing quickly
   * coding population genetic simulation models can be arduous and error-prone
     + stdpopsim is a tool recently developed to provide easy access to simulation frameworks
       - avoid re-making common models
       - provide standard benchmarks for methods development and testing
       - for both methods development and empirical research
     + but so far stdpopsim has been mainly restricted to well-characterised model organisms, limiting utility for non-model organisms or model organisms in the early stages of development
   * many potential users want to simulate their study organism, or develop inference methods for non-model organisms
     + feedback from 2020/2021 stdpopsim workshops emphasized the need for simulations of a wide variety of model and non-model species
2. **ELISE** Therefore this paper is intended as a resource for both methods developers and empirical researchers to develop simulations of their own species of interest, with the potential to submit the simulation framework for inclusion in the stdpopsim catalog
   * we discuss the elements of a population genomic simulation model that realistically characterizes a species, including:
     + input data
     + reasonable data quality
     + simulation frameworks (coalescent and forward?)
   * in addition, we discuss
     + limitations of available data and when a species-specific simulation may be impractical or unnecessary for the intended task
   * lastly, we discuss how these models may be integrated into the stdpopsim catalog
     + we present the current method for adding species and demographic scenarios to the catalog
     + and clarify the required genomic resources and QC process

## Tutorial

1. Making a popgen simulation

- necessary resources:  
 - genome assembly  
 - this needs to be chromosome level, or nearly so  
 - contigs are not sufficient because:  
 - lack of linkage information limits their use for inference  
 - (see below for further discussion of species with contig-level assemblies)  
 - a plausible, citeable mutation rate  
 - recombination map - or at least a plausible, citeable, recombination rate  
 - a citeable demographic model, or at least a plausible, citeable, effective population size  
 - annotations (required if the use case involves coding/noncoding regions, specific genes, etc.)  
- draw from https://github.com/popsim-consortium/workshops/blob/main/adding\_species/contributing.ipynb without specifics for stdpopsim   
- emphasize throughout that these features should be citeable and chosen with deliberation  
 - eg. "often reported parameter values are tied to a model and its other parameters, so mixing values from different sources has its own caveats" (https://github.com/popsim-consortium/adding-species-manuscript/issues/7)  
- what about species that don't have some of those resources?  
 - if everything is available except the genome is contig-level (and no recombination map):  
 - would want to choose just a few of the most important or longest contigs to simulate  
 - https://github.com/popsim-consortium/adding-species-manuscript/issues/1#issuecomment-1050081855  
 - might be more appropriate to just simulate a locus of appropriate length

Table of “what if one of the features above required are missing”

| missing paramet | er | options | considerations |
| --- | --- | --- |
| mutation rate | borrow from c | losest relative with citeable rate? | will affect estimates of (timing in years, …) |
| recombination r | ate | borrow f | rom closest relative with citeable rate? | will affect patterns of (selection/linkage, …) |
| a good assembly consider just s | | select the imulating a lo | largest/most contiguous contigs, and/or those with genes of interest if known +——————–+——————–+ cus of appropriate length if you don’t somehow have a recombination map or a gene(s) or other loci of interest with known contig location |
| demographic mod | el | need at l | east Ne | bottlenecks/expansions/admixture can affect a lot of patterns of variation |
| Ne | population ge | nomic simulation not practical? | (I’m not sure there’s much you can do here except pick something vaguely reasonable and pray) |

## Application/Discussion

1. **ELISE** Lessons from “Growing the Zoo” hackathon, held along side probgen in April 2021
   * Community-based expansion of the number and variety of species and their demographic scenarios included in the stdpopsim catalog
     + when first published, the stdpopsim catalog included 6 species: *Homo sapiens*, *Pongo abelii*, *Canis familiaris*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Escherichia coli.*
     + when did additional demographic scenarios start getting added?
     + the catalog now includes an additional **12** species (note: is this correct, number is from those on main branch)? the catalog doc still only includes the original 6? - MEL), as well as multiple demographic scenarios for *Homo sapiens*, *Pongo abelii*, *Drosophila melanogaster*, and *Arabidopsis thaliana.* (others?)
   * We had *X* participants in the introduction to stdpopsim workshops, many of whom expressed a wish to add their own species for empirical work
     + further *Y* participants in the hackathon
     + stdpopsim has become a popular resource and there is clearly a desire/need for adding additional species for both goals of methods development and empirical inference
     + Only *Z* species ultimately got added, why?
       - choosing species/demographic scenarios to add is potentially challenging
       - for methods development, they have to be appropriate for the goals of the method
       - for empirical studies, the species is obvious, but whether it is appropriate for inclusion in the catalog or even to use stdpopsim may be less clear
     + Current gaps in types of species in catalog
       - 7 chordates, 2 plants (1 vascular, 1 algae), 2 bacteria, 6 arthropods, 1 nematode
       - lots of potential for supporting methods development for other types of organisms (that aren’t vertebrates and insects)
         * eg. yeast
         * eg. more plants because plants have goofy genomes
2. Examples from “Zoo” hackathon
   * A species that was fully added (Bos taurus?)
     + resources available
     + reasons for adding
     + can be used for methods development and inference
   * A demographic model that was fully added (note: not sure there were any? we worked on Canis familiaris but it hasn’t finished QC - MEL)
     + resources available (already had species in catalog, but not a demographic model, just Ne)
     + reasons for adding
   * A species that was purposefully not added (Myotis lucifugus)
     + resources available (everything except chromosome-level genome)
     + why wasn’t it added? (see above re: contigs and just using msprime directly)

Table of species/demographic scenarios added/worked on by community since original pub

| Species | Phylum | New species or demog. model? | Outcome??? (added, abandoned, in QC, chosen not to add?) |
| --- | --- | --- | --- |
| Bos taurus | Chordata | Both | Added |
| … | … | … | … |

1. Steps to add a species to stdpopsim catalog
   * find required resources and citations (chromosome-level assembly, mutation rate, recombination rate or map, Ne or demog. model, annotations) - including an actual human being(s) willing to put in the time and effort - find and evaluate the necessary information and citations - add code - detailed steps, reference workshop materials on github <https://github.com/popsim-consortium/workshops/blob/main/adding_species/contributing.ipynb>
   * QC - detailed steps
   * update catalog documentation?
   * emphasize that anything that ends up in the catalog has the potential to be updated (because science)

## Conclusion

1. Take-aways
   * we present the basics for
     + 1. determining if a species-specific population genomic simulation is appropriate for the species/question
       2. getting the required data
       3. creating a population genomic simulation of that species in your simulator of choice
       4. adding that species to the stdpopsim catalog
     + we *do not* present tutorials for implementation in any particular simulator (there are lots of those)
   * there is a definite need for more and more species-specific population genomic simulations
     + but gaps in practitioners’ understandings of when and how to implement them for a given species
     + stdpopsim is a community-maintained resource intended to provide easy access to simulation frameworks, adding species expands the types of methods development and inference it is useful for
       - we learned from the hackathon that people are willing to put in the time and effort to add species, when given clear guidance for how to do it
       - not all species that people want to add have appropriate genomic resources yet, but many will soon!
       - so this paper provides the population genomics community with more clarity in what resources are necessary and how to use them to simulate your favorite species

# Meeting notes:

* recap goals and structure of stdpopsim
  + two sets of goals:
    - methods development/testing
    - inference (this is perhaps the major goal this paper is addressing?)
    - arguably adding species opens stdpopsim to the community whose primary concern is inference rather than methods development
* discussion of resources necessary to simulate/add species
  + recombination map
  + annotation
  + assembly
    - not just contigs
  + demographic model?
    - at least reasonable Ne
* discussion of types of species in terms of:
  + available resources
  + what goal adding the species would address
* description of how to add a species to the catalog and/or write a yaml for a new species
  + this will depend on how we’re ultimately moving forward - using the current python-based method, or switching to a yaml-based method