simuPOP Reference Manual

Release 0.8.0 (Rev: 1209)

Bo Peng

December 2004

Last modified
14th August 2007

Department of Epidemiology, U.T. M.D. Anderson Cancer Center
Email: bpeng@mdanderson.org
URL: http://simupop.sourceforge.net
Mailing List: simupop-list@lists.sourceforge.net

Acknowledgements:

Dr. Marek Kimmel
Dr. François Balloux
Dr. William Amos
SWIG user community
Python user community
Keck Center for Computational and Structural Biology
U.T. M.D. Anderson Cancer Center

Abstract

simuPOP is a forward-time population genetics simulation environment. Unlike coalescent-based programs, simuPOP evolves populations forward in time, subject to arbitrary number of genetic and environmental forces such as mutation, recombination, migration and population/subpopulation size changes. Statistics of populations can be calculated and visualized dynamically which makes simuPOP an ideal tool to demonstrate population genetics models; generate datasets under various evolutionary settings, and more importantly, study complex evolutionary processes and evaluate gene mapping methods.

The core of simuPOP is a scripting language (Python) that provides a large number of building blocks (populations, mating schemes, various genetic forces in the form of operators, simulators and gene mapping methods) to construct a simulation. This provides a R/Splus or Matlab-like environment where users can interactively create, manipulate and evolve populations, monitor and visualize population statistics and apply gene mapping methods. The full power of simuPOP and Python (even R) can be utilized to simulate arbitrarily complex evolutionary scenarios.

simuPOP is written in C++ and is provided as Python modules. Besides a front-end providing an interactive shell and a scripting language, Python is used extensively to pass dynamic parameters, calculate complex statistics and write operators. Because of the openness of simuPOP and Python, users can make use of external programs, such as R, to perform statistical analysis, gene mapping and visualization. Depending on machine configuration, simuPOP can simulate large (think of millions) populations at reasonable speed.

This is a reference manual to all variables, functions, and objects of simuPOP. To learn different components of simuPOP and how to write simuPOP scripts, please refer to the *simuPOP User's Guide.*

How to cite simuPOP:

> Bo Peng and Marek Kimmel (2005) simuPOP: a forward-time population genetics simulation environment. *bioinformatics*, **21**(18): 3686-3687

CONTENTS

LIST OF EXAMPLES

CHAPTER

ONE

Introduction

This reference manual assumes that you have read the *simuPOP User's Guide* and know the basic concepts of simuPOP. It is also remcommended that you learn some basics of Python before you continue. I have listed a few Python resources at the end of this chapter, along with links to some simuPOP tutorials.

Almost all information contained in this manual can be accessed from command line, after you install and import the simuPOP module. For example, you can use `help(population.addInfoField)` to view the help information of member function `addInfoField` of class `population`.

Example 1.1: Getting help using the help() function

```
>>> help(population.addInfoField)
Help on method population_addInfoField:

population_addInfoField(...) unbound simuPOP_la.population method
    Description:

        add an information field to a  population

    Usage:

        x.addInfoField(field, init=0)

    Arguments:

        field:          new information field. If it already exists, it
                        will be re-initialized.
        init:           initial value for the new field.

>>>
```

It is important that you understand that

- The constructor of a class is named `__init__` in python. That is to say, you should use the following command to display the help information of the constructor of class `population`:

  ```
  >>> help(population.__init__)
  ```

- Some classes are derived from other classes and have access to member functions of their base classes. For example, class `population`, `individual` and `simulator` are all derived from class `GenoStruTrait`. Therefore, you can use all `GenoStruTrait` member functions from these classes.

  The constructor of a derived class also calls the constructor of its base class so you may have to refer to the base class for some parameter definitions. For example, parameters `begin, end, step, at` etc are shared by all operators, and are explained in details only in class `baseOperator.`

1.1 Loading simuPOP

simuPOP is composed of six modules: standard short, long and binary alleles, each of them have standard and optimized modules. A Message Passing Interface (MPI) version is under development but not yet available. The short modules use 1 byte to store each allele which limits the possible allele states to 256. This is enough most of the times but not so if you need to simulate models such as the infinite allele model. In those cases, you can use the long allele version of the modules, which use 2 bytes for each allele and can have 2^{16} possible allele states. On the other hand, if you would like to simulate a large number of binary (SNP) markers, binary modules can save you a lot of RAM. Depending on applications, binary modules can be faster or slower than other modules.

Standard modules have detailed debug and run-time validation mechanism to make sure the simulations run correctly. Whenever something unusual is detected, simuPOP would terminate with a detailed error message. The cost of such run-time checking varies from application to application but can be very high under some extreme circumstances. Because of this, optimized versions for all modules are provided. They bypass all parameter checking and run-time validations and will simply crash if things go wrong. It is recommended that you use standard modules whenever possible and only use the optimized version when performance is needed and you are confident that your simulation is running as expected.

Example 1.2 and 1.3 demonstrate the differences between standard and optimized modules, by executing two invalid commands. The standard module returns proper error messages, while the optimized module returns erroneous results and even crashes.

Example 1.2: Use of standard simuPOP modules

```
>>> pop = population(10, loci=[2])
>>> pop.locusPos(10)
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
IndexError: src/genoStru.h:447 absolute locus index (10) out of range of 0 - 1
>>> pop.individual(20).setAllele(1, 0)
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
IndexError: src/population.h:443 individual index (20) is out of range of 0 ~ 9
>>>
```

Example 1.3: Use of optimized simuPOP modules

```
% setenv SIMUOPTIMIZED
% python
>>> from simuPOP import *
simuPOP : Copyright (c) 2004-2006 Bo Peng
Developmental Version (May 21 2007) for Python 2.3.4
[GCC 3.4.6 20060404 (Red Hat 3.4.6-3)]
Random Number Generator is set to mt19937 with random seed 0x2f04b9dc5ca0fc00
This is the optimied short allele version with 256 maximum allelic states.
For more information, please visit http://simupop.sourceforge.net,
or email simupop-list@lists.sourceforge.net (subscription required).
>>> pop = population(10, loci=[2])
>>> pop.locusPos(10)
1.2731974748756028e-313
>>> pop.individual(20).setAllele(1, 0)
Segmentation fault
```

You can control the choice of modules in the following ways:

- Set environment variable `SIMUALLELETYPE` to 'short', 'long' or 'binary', and `SIMUOPTIMIZED` to use the optimized version. The default module is the standard short module.
- Before you load simuPOP, set options using `simuOpt.setOptions(optimized, alleleType, quiet, debug)`. `alleleType` can be `short, long` or `binary. quiet=True` suppresses banner information when simuPOP is loaded, and `debug` is a comma-separated list of debug options specified by `listDebugCode`(). Debug information is only available for standard modules.
- If you are running a simuPOP script that conforms to simuPOP convention, you should be able to use optimized module using command line option `--optimized.`

After a simuPOP module is loaded, you can use the following functions to determine some module and platform dependent information.

- `AlleleType()`: return `'binary'`, `'short'`, or `'long'`.
- `Optimized()`: return `True` or `False`.
- `MaxAllele`(): return `1` for binary modules, usually `255` for short modules and $2^{16}-1$ for long modules.
- `simuVer()`: return the version string
- `simuRev()`: simuPOP revision number. If your script needs a recent version of simuPOP, it is a good idea to test `simuRev()` against the revision when the feature you need becomes available.
- `Limits()`: print the limits of this module on this platform, such as the maximum number of loci in a population.

Example 1.4: Use simuOpt to control which simuPOP module to load

```
>>> import simuOpt
>>> simuOpt.setOptions(optimized=False, alleleType='long', quiet=True)
>>> from simuPOP import *
>>> print alleleType()
long
>>> print optimized()
False
>>>
```

1.2 References and the `clone()` member function

Assignment in Python only creates a new refernce to the exsting object. For example,

```
pop = population(...)
pop1 = pop
```

will create a reference `pop1` to population `pop`. Modifying `pop1` will modify `pop` as well. If you would like to have an independent copy, use

```
pop1 = pop.clone()
```

All simuPOP classes (objects) have a clone function that can be used to create an independent copy of the object. Because cloning a large population can be costly, a few methods are provided to access populations inside a simulator. Assuming that `simu` is a simulator with several populations,

1. `simu.population(rep)` returns a reference to the `rep`'th population. You can, although not recommended, modify simulator through this `pop` reference. Be cautious though, that the following seemingly innocent usage of this function will crash simuPOP, because the simulator `simu` will be destroyed after the call to `func()` is ended, leaving `pop` as a referece to an invalid population object.

   Example 1.5: Reference to a population of a simulator

   ```
   def func():
       simu = simulator(
           population(10),
           randomMating())
       # evolve simu ..., then return population
       return simu.population(0)

   pop = func()
   pop.popSize()
   ```

2. To get an independent copy of a population, you can use `pop = simu.getPopulation(rep)`, which returns an independent copy of population `rep` of `simu`. `simu` is untouched.

3. If the simulator will be destroyed as in Example 1.5,

   ```
   pop = simu.getPopulation(rep, destructive=True)
   ```

   can be used. This function will *extract* population `rep` from the simulator instead of copying it, and bypassing a potentially very costly process.

1.3 Zero-based indexes, ranges, absolute and relative indexes

All arrays in simuPOP start at index 0. This conforms to Python and C++ indexes. To avoid confusion, I will refer the first locus as locus zero, the second locus as locus one; the first individual in a population as individual zero, and so on.

Ranges in simuPOP also conforms to Python ranges. That is to say, a range has the form of `[a,b)` where `a` belongs to the range, and `b` does not. For example, `pop.chromBegin(1)` refers to the index of the first locus on chromosome 1 (actually exists), and `pop.chromEnd(1)` is the index of the last locus on chromosome 1 **plus 1**, which might or might not be a valid index. In this way

```
for locus in range(pop.chromBegin(1), pop.chromEnd(1)):
    print locus
```

will iterate through all locus on chromosome 1.

Another two important concepts are the *absolute index* and the *relative index* of a locus. The former index ignores chromosome structure. For example, if there are 5 and 7 loci on the first two chromosomes, the absolute indexes of the two chromosomes are (0, 1, 2, 3, 4), (5, 6, 7, 8, 9, 10, 11) and the relative indexes are (0, 1, 2, 3, 4), (0, 1, 2, 3, 4, 5, 6). Absolute indexes are more frequently used because they avoid the trouble of having to use two numbers (chrom, index) to refer to a locus. Two functions `chromLocusPair(absIndex)` and `absLocusIndex(chrom,index)` are provided to convert between these two kinds of indexes. An individual can also be referred by its *absolute index* and *relative index* where *relative index* is the index in its subpopulation.

Example 1.6: Conversion between absolute and relative indices

```
>>> pop = population(subPop=[20, 30], loci=[5, 6])
>>> print pop.chromLocusPair(7)
(1, 2)
>>> print pop.absLocusIndex(1,1)
```

```
6
>>> print pop.absIndIndex(10, 1)
30
>>> print pop.subPopIndPair(40)
(1, 20)
>>>
```

1.4 Function form of an operator

Operators are usually applied to populations through a simulator. They are created and passed as parameters to the `evolve` function of a simulator. During evolution, the `evolve()` function determines if an operator can be applied to a population and apply it when appropriate. More details about operators will be described in section **??**.

You can ignore the specialties of an operator and call its `apply()` function directly. For example, you can initialize a population outside a simulator by

```
initByFreq( [.3, .2, .5] ).apply(pop)
```

or dump the content of a population by

```
dumper().apply(pop)
```

This usage is used so often that it deserves some simplification. Equivalent functions are defined for most of the operators. For example, function `InitByFreq` is defined for operator `initByFreq` as follows

Example 1.7: Function InitByFreq

```
>>> def InitByFreq(pop, *args, **kwargs):
...     initByFreq(*args, **kwargs).apply(pop)
...
>>> InitByFreq(pop, [.2, .3, .4, .1])
>>>
```

The function form of an operator is listed after its class name in this reference manual.

1.5 The `carray` type

The return value of simuPOP functions with names starting with `arr` is of a special Python type `carray`. This object reflects the underlying C/C++ array which can be modified through this list-like interface, with the exception that you can not change the size of the array. Only `count` and `index` list functions can be used, but all comparison, assignment and slice operations are allowed.

Example 1.8: Usage of the carray type

```
>>> # obtain an object using one of the arrXXX functions
>>> pop = population(loci=[3,4], lociPos=[1,2,3,4,5,6,7])
>>> arr = pop.arrLociPos()
>>> # print and expression (just like list)
>>> print arr
[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0]
>>> str(arr)
'[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0]'
>>> # count
>>> arr.count(2)
```

```
1
>>> # index
>>> arr.index(2)
1
>>> # can read write
>>> arr[0] = 0.5
>>> # the underlying locus position is also changed
>>> print pop.lociPos()
(0.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0)
>>> # convert to list
>>> arr.tolist()
[0.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0]
>>> # or simply
>>> list(arr)
[0.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0]
>>> # compare to list directly
>>> arr == [0.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0]
True
>>> # you can also convert and compare
>>> list(arr) == [0.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0]
True
>>> # slice
>>> arr[:] = [1,2,3,4,5,6,7]
>>> # assign from another part
>>> arr[1:3] = arr[3:5]
>>> # arr1 is 1,2,3
>>> arr1 = arr[:3]
>>> # assign slice from a number
>>> # arr will also be affected since arr1 point to a part of arr
>>> arr1[:] = 10
>>> # assign vector of the same length
>>> len(arr1)
3
>>> arr1[:] = [30,40, 50]
>>>
```

Important note: Objects returned from `arrXXX` functions should be considered temporary. There is no guarantee that the underlying array will still be valid after any population operation.

1.6 Random Number Generator

There are many random number generators (RNG) with different properties. Using a bad RNG can serious compromise the validity of simulation results. Although the default RNG `mt19937` has good performance, simuPOP allows you to choose from a number of RNGs, all from GNU Scientific Library (GSL). Please refer to the documentation of GSL for more details about these RNGs.

If you need to use a RNG in your simuPOP script, you can either use Python random module (`import random`) , use `rng()` function to get the RNG of simuPOP, or create a separate RNG using `RNG(name, seed)` function.

Example 1.9: Random number generator

```
>>> print ListAllRNG()
('gfsr4', 'mt19937', 'mt19937_1999', 'mt19937_1998', 'r250', 'rand', 'rand48', 'random12
>>> print rng().name()
```

```
mt19937
>>> SetRNG("taus2", seed=10)
>>> print rng().name()
taus2
>>>
```

1.6.1 Random seed

When simuPOP is loaded, it creates a random number generator. This RNG gets its seed from:

- random number from `/dev/urandom` if it is available
- random number from `/dev/random` if it is available
- use Python `random` module and assign random seed with

```
stlisting

\end{itemize}
The last method is used mostly under windows, on which random devices
are not available. It tries to avoid sole use of system time because
simultaneously started jobs may get the same random seed.

Random seed can be retrieved using \texttt{rng().seed()} function,
which can be saved for future reference.

\RNGRef


\section{Name Conventions}

\texttt{simuPOP} follows the following naming conventions.

\begin{itemize}
\item Classes (objects), member functions and parameter names start with
small character and use capital character for the first character
of each word afterward. For example
\end{itemize}
\begin{lyxcode}
population,~population::subPopSize(),~individual::setInfo()
\end{lyxcode}
\begin{itemize}
\item Most standalone (global) functions start with capital character. This
is how you can differ an operator from its function version. For example,
\texttt{initByFreq(vars)} is an operator and \texttt{InitByFreq(pop,
vars)} is its function version.
\item Constants start with capital characters. For example
\end{itemize}
\begin{lyxcode}
MigrByProportion,~StatNumOfFemale
\end{lyxcode}
\begin{itemize}
```

```
\item The following words in function names are abbreviated:
\end{itemize}
\begin{lyxcode}
pos~(position),~~info~(information),~migr~(migration),~subPop~(subpopulation),~~\\
(rep)~replicate,~gen~(generation),~grp~(group(s)),~ops~(operators),~~\\
expr~(expression),~stmts~(statements)~
\end{lyxcode}

\section{Online resources}

There are several excellent Python books and tutorials. If you are
new to Python, you can start with

\begin{enumerate}
\item The Python tutorial (\texttt{http://docs.python.org/tut/tut.html})
\item Other online tutorials listed at http://www.python.org/doc/
\end{enumerate}
The PDF version of this reference manual is distributed with simuPOP.
You can also get the latest version of this file from the simuPOP
subversion repository. To access it, go to \texttt{http://simupop.sourceforge.net},
click\textsf{ SF.net summary > Code > SVN Browse > trunk > doc > refManual.pdf}
and download the HEAD version. You can also find some tutorials that
are not distributed with simuPOP from the subversion repository, such
as

\begin{enumerate}
\item Forward-time simulations using simuPOP, a tutorial: a tutorial that
was given in a simuPOP workshop held at University of Alabama at Birmingham.
\item Forward-time simulations using simuPOP, an in-depth course: a in-depth
course about simuPOP components, with a lot of examples.
\end{enumerate}
The filenames are \texttt{tutorial.pdf }and \texttt{course.pdf}, respectively.
Note that these presentations will not be updated so their content
can become out of date. This reference manual should be considered
as the authoritative resource of simuPOP.


\chapter{\,simuPOP Components}


\section{Genotypic structure \index{genotypic structure}}

Genotypic structure refers to

\begin{itemize}
\item ploidy, the number of copies of basic number of chromosomes (c.f.
\texttt{ploidy\index{GenoStruTrait!ploidy}(), ploidyName\index{GenoStruTrait!ploidyN
)
\item the number of chromosomes (c.f. \texttt{numChrom\index{GenoStruTrait!numChrom}
)
\item the existence of sex chromosome (c.f. \texttt{sexChrom\index{GenoStruTrait!sex
)
\item the number of loci on each chromosome (c.f. \texttt{numLoci\index{GenoStruTrai
totNumLoci\index{GenoStruTrait!totNumLoci}()} )
```

\item the locus position on its chromosome (c.f. \texttt{locusPos\index{GenoStruTrai
arrLociPos\index{GenoStruTrait!arrLociPos}()})
\item allele name(s), default to allele number (c.f. \texttt{alleleName\index{GenoSt
, \texttt{alleleNames()})
\item the maximum allele state (c.f. \texttt{maxAllele\index{GenoStruTrait!maxAllele
)
\item the names of the information fields (c.f. \texttt{infoField\index{GenoStruTrai
infoFields\index{GenoStruTrait!infoFields}()})
\end{itemize}
\emph{Information fields} refer to float numbers attached to each
individual, such as fitness value, parent index, age. They are used
to store auxiliary information of individuals, **and** are essential to
the operations of some simuPOP components. For example, \texttt{'fitness'}
field **is** required by all selectors. Details please refer to section
\ref{sec:Information-fields}.

If \texttt{sexChrom()} **is** false, all chromosomes are assumed to be
autosomes. You can also create populations with a sex chromosome.
Currently, simuPOP only models the XY chromosomes **in** diploid population.
This **is** to say,

\begin{itemize}
\item sex chromosome **is** always the last chromosome.
\item sex chromosome can only be specified **for** diploid population (\texttt{ploidy()=
\item sex chromosomes (XY) may differ **in** length. You should specify the
length of the longer one as the chromosome length. If there are more
loci on X than Y, the rest of the Y chromosome **is** unused. Mutation
may still occur at this unused part of chromosome to simplify implementation
and usage.
\item it **is** assumed that males have XY **and** females have XX chromosomes.
The sex chromosomes of male individuals are **in** the order of XY.
\end{itemize}
Individuals **in** the same population share the same genotypic structure.
Consequently, \emph{the genotypic information can be accessed **from**
individual, population **and** simulator} \emph{levels}.

\GenoStruTraitRef

\section{Population}

\texttt{population\index{population}} objects are essential to simuPOP.
They are composed of subpopulations each with certain number of individuals
having the same genotypic structure. Class \texttt{population} has
a large number of member functions, ranging **from** reviewing simple
properties to generating a new population **from** the current one. Fortunately,
you do **not** have to know all the member functions to use a population
unless you need to write pure Python functions/operators that involves
complicated manipulation of populations.

simuPOP uses one-level population structure. That **is** to say, there
is no sub-subpopulation **or** family **in** subpopulations. Mating **is** within
subpopulations only. Exchanges of genetic information across subpopulations
can only be done through migration. Population **and** subpopulation sizes

```
can be changed, as a result of mating or migration. More specifically,

\begin{itemize}
\item migration can change subpopulation size; create or remove subpopulations.
Since migration can not generate new individuals, the total population
size will not be changed.
\item mating can fill any population/subpopulation structure with offspring.
Both population and subpopulation sizes can be changed. Since mating
is within subpopulations, you can not create a new subpopulation through
mating.
\item a special operator \texttt{pySubset} can shrink the population size.
It removes individuals according to their \texttt{subPopID()} status.
(Will explain later.) This can be used to model a sudden population
decrease due to some natural disaster.
\item subpopulations can be split or merged.
\end{itemize}
Note that migration will most likely change the subpopulation sizes.
To keep the subpopulation sizes constant, you can set the subpopulation
sizes during mating so that the next generation will have desired
subpopulation sizes.

\populationRef


\subsection{\label{sub:Ancestral-populations}Ancestral populations}

By default, a population object only holds the current generation.
All ancestral populations (generations) will be discarded. You can,
however, keep as many ancestral generations as you wish, provided
that you have enough RAM to store all these extra information.

Parameter \texttt{ancestralDepth} is used to specify the number of
generations to keep. This parameter is default to \texttt{0}, meaning
keeping no ancestral population. You can specify a positive number
\texttt{n} to store most recent n generations; or -\texttt{1} to store
all populations.

Several important usage of ancestral generations:

\begin{itemize}
\item \texttt{dumper()} operator and \texttt{Dump()} function has a parameter
\texttt{ancestralPops}. If set to \texttt{True}, they will dump all
ancestral generations.
\item function \texttt{population::setAncestralDepth()} and operator \texttt{setAnce
set the number of ancestral generations to keep for a population.
A typical use of \texttt{setAncestralDepth()} is



\begingroup
\inputencoding{latin1}
\begin{lstlisting}
simu.evolve(...
  setAncestralDepth(3, at=[-3])
```

```
    )
```

which saves the last three generations in populations so that pedigree based sampling schemes can be used.

- `pop.useAncestralPop(idx)` set the current generation of population `pop` to `idx` generation. `idx = 1` for the first ancestral generation, 2 for second ancestral ..., and 0 for the current generation. After this function, all functions, operators will be applied to this ancestral generation. You should always call `setAncestralPop(0)` after you examined the ancestral generations.

A typical use of this function is demonstrated in example **??**. In this example, a population with two loci is created and with initial genotype 0. Two `kamMutator` with different mutation rates are applied to these two loci. Five most recent populations are kept. The allele frequencies at these generations are calculated afterward. (Note that this is not the best way to exam the changes of allele frequencies, a `stat` operator should be used.)

Example 1.10: Ancestral populations

```
>>> simu = simulator(population(10000, loci=[2]), randomMating())
>>> simu.evolve(
...     ops = [
...         setAncestralDepth(5, at=[-5]),
...         kamMutator(rate=0.01, loci=[0], maxAllele=1),
...         kamMutator(rate=0.001, loci=[1], maxAllele=1)
...     ],
...     end = 20
... )
True
>>> pop = simu.population(0)
>>> # start from current generation
>>> for i in range(pop.ancestralDepth()+1):
...     pop.useAncestralPop(i)
...     Stat(pop, alleleFreq=[0,1])
...     print '%d     %5f     %5f' % \
...         (i, pop.dvars().alleleFreq[0][1], pop.dvars().alleleFreq[1][1])
...
0     0.166850     0.015200
1     0.162550     0.014300
2     0.158600     0.013900
3     0.148050     0.013350
4     0.140650     0.013550
5     0.133950     0.012650
>>> # restore to the current generation
>>> pop.useAncestralPop(0)
>>>
```

1.6.2 Save and Load a Population

Internally, population can be saved/loaded in "`txt`", "`xml`" or "`bin`" formats using `savePopulation(file, format, compress=True)` member function, global `SavePopulation(pop, file, format)` and `LoadPopulation`. (Yes, it is `Load..` not `load..` since `savePopulation` is a member function and `LoadPopulation` is a global function.) These formats have their own advantages and disadvantages:

- `xml:` most readable, easy transformation to other formats, largest file size

- `bin:` not readable, small file size. May not be portable.
- `txt:` human readable with no structure, portable, median file size.
- `auto :` the format is determined by the filename extension specified.

Populations are by default compressed in gzip format. If you are interested in viewing the content of the file, you can use `compress=False` when saving a population, or decompress the saved files using `gzip -d` command.

Populations can also be saved in other formats such as FSTAT so that they can be directly analyzed by other programs. These formats are not supported internally. They are handled in Python in the form of Python function or pure-Python operator. If you would like to save/load simuPOP population in your own format, you can do it by mimicking these functions in `simuUtil.py`.

Shared variables (c.f section **??**) are also saved (except for big objects like samples). Since the number of shared variables can be very large, it maybe a good idea to clear these variables before you save a population. On the other hand, you may want to save key parameters used to generate this population in the local namespace so that you will know these parameters after the population is loaded. For example, you can do

Example 1.11: Save population variables

```
pop.vars().clear()
pop.dvars().migrationRate = 0.002
pop.dvars().diseaseLoci = [4, 30]
SavePopulation(pop, 'pop.bin')
```

1.6.3 View a population (GUI, wxPython required)

Introduced in version 0.6.9, `simuViewPop.py` can be used to view a population. It can be used as a standalone application, or in an interactive session. First, you can use this script as a standalone application, simply run

```
simuViewPop.py mypop.bin
```

will fire a GUI and allow you to exam population property, genotype and calculate statistics.

In a Python session, import this module will provide a function `viewPop`, apply it on a in-memory population or a filename will have the same effect. For example,

Example 1.12: Use simuViewPop to view a population

```
import simuViewPop
simuViewPop.viewPop(myPop)
simuViewPop.viewPop(filename='mypop.bin')
```

1.7 Individuals

Individuals of a population can be accessed through `individual()`, or its iteration form `individuals()` function:

- `individual(ind)` returns the `ind`'th individual (absolute index) of the whole population.
- `individual(ind, subPop)` returns the `ind`'th (relative index) individual in the `subPop`'th subpopulation.
- `individuals()`return an iterator that can be used to iterate through all individuals in a population.

- `individuals(subPop)` return an iterator that can be used to iterate through all individuals in the `subPop`'th subpopulations.

For example, example **??** iterates through all individuals in subpopulaiton 2 using `population::individual()` function, while **??** uses `population::individuals()`. The latter is usually easier to use.

Example 1.13: Function population::individual()

```
for i in range(pop.subPopSize(2)):
    ind = pop.individual(i, 2)
    print ind.affected()
```

Example 1.14: Function population::individuals()

```
for ind in pop.individuals(2):
    # do something to ind
    print ind.affected()
```

1.7.1 Class `individual`

Individuals with genotype, affection status, sex etc.

Details

Individuals are the building blocks of populations, each having the following individual information:

- shared genotypic structure information
- genotype
- sex, affection status, subpopulation ID
- optional information fields

Individual genotypes are arranged by locus, chromosome, ploidy, in that order, and can be accessed from a single index. For example, for a diploid individual with two loci on the first chromosome, one locus on the second, its genotype is arranged as `1-1-1 1-1-2 1-2-1 2-1-1 2-1-2 2-2-1` where `x-y-z` represents ploidy `x` chromosome `y` and locus `z`. An allele `2-1-2` can be accessed by `allele(4)` (by absolute index), `allele(1, 1)` (by index and ploidy) or `allele(1, 1, 0)` (by index, ploidy and chromosome).

Initialization

Individuals are created by populations automatically. Do not call this function directly.

```
individual()
```

Member Functions

`x.affected()` Whether or not an individual is affected

`x.affectedChar()` Return `A` or `U` for affection status

`x.allele(index)` Return the allele at locus `index`

index absolute index from the beginning of the genotype, ranging from `0` to `totNumLoci()*ploidy()`

`x.allele(index, p)` Return the allele at locus `index` of the `p-th` copy of the chromosomes

index index from the begining of the `p-th` set of the chromosomes, ranging from `0` to `totNumLoci()`

p index of the ploidy

`x.allele(index, p, ch)` Return the allele at locus `index` of the `ch-th` chromosome of the `p-th` chromosome set

ch index of the chromosome in the `p-th` chromosome set

index index from the begining of chromosome `ch` of ploidy `p`, ranging from `0` to `numLoci(ch)`

p index of the polidy

`x.alleleChar(index)` Return the name of `allele(index)`

`x.alleleChar(index, p)` Return the name of `allele(index, p)`

`x.alleleChar(index, p, ch)` Return the name of `allele(idx, p, ch)`

`x.arrGenotype()` Return an editable array (a carray of length `totNumLoci()*ploidy()`) of genotypes of an individual

This function returns the whole genotype. Although this function is not as easy to use as other functions that access alleles, it is the fastest one since you can read/write genotype directly.

`x.arrGenotype(p)` Return a carray with the genotype of the `p-th` copy of the chromosomes

`x.arrGenotype(p, ch)` Return a carray with the genotype of the `ch-th` chromosome of the `p-th` copy

`x.arrInfo()` Return a carray of all information fields (of size `infosSize()`) of this individual

`x.info(idx)` Get information field `idx`

idx index of the information field

`x.info(name)` Get information field `name`

Equivalent to `info(infoIdx(name))`.

name name of the information field

`x.setAffected(affected)` Set affection status

`x.setAllele(allele, index)` Set the allele at locus `index`

allele allele to be set

index index from the begining of genotype, ranging from `0` to `totNumLoci()*ploidy()`

`x.setAllele(allele, index, p)` Set the allele at locus `index` of the `p-th` copy of the chromosomes

allele allele to be set

index index from the begining of the poloidy `p`, ranging from `0` to `totNumLoci(p)`

p index of the poloidy

`x.setAllele(allele, index, p, ch)` Set the allele at locus `index` of the `ch-th` chromosome in the `p-th` chromosome set

allele allele to be set

ch index of the chromosome in ploidy `p`

index index from the begining of the chromosome, ranging from `0` to `numLoci(ch)`

p index of the ploidy

`x.setInfo(value, idx)` Set information field by `idx`

x.setInfo(value, name) Set information field by `name`

x.setSex(sex) Set sex. `sex` can be `Male` of `Female`.

x.setSubPopID(id) Set new subpopulation ID, `pop.rearrangeByIndID` will move this individual to that population

x.sex() Return the sex of an individual, `1` for males and `2` for females.

x.sexChar() Return the sex of an individual, `M` or `F`

x.subPopID() Return the ID of the subpopulation to which this individual blongs

Note: `subPopID` is not set by default. It only corresponds to the subpopulation in which this individual resides after `pop::setIndSubPopID` is called.

x.unaffected() Equals to `not affected()`

Example

Example 1.15: Individual member functions

```
>>> pop = population(500, loci=[2, 5, 10])
>>> # get an individual
>>> ind = pop.individual(9)
>>> # oops, wrong index
>>> ind = pop.individual(3)
>>> # you can access genotypic structure info
>>> print ind.ploidy()
2
>>> print ind.numChrom()
3
>>> # ...
>>> # as well as genotype
>>> print ind.allele(1)
0
>>> ind.setAllele(1,5)
>>> print ind.allele(1)
0
>>> # you can also use an overloaded function
>>> # with a second parameter being the ploidy index
>>> print ind.allele(1,1) # second locus at the second copy of chromosome
0
>>> # other information
>>> print ind.affected()
False
>>> print ind.affectedChar()
U
>>> ind.setAffected(1)
>>> print ind.affectedChar()
A
>>> print ind.sexChar()
M
>>>
```

1.8 Mating Scheme

1.8.1 Class `mating`

The base class of all mating schemes - a required parameter of `simulator`

Details

Mating schemes specify how to generate offspring from the current population. It must be provided when a simulator is created. Mating can perform the following tasks:

- change population/subpopulation sizes;
- randomly select parent(s) to generate offspring to populate the offspring generation;
- apply *during-mating* operators;
- apply selection if applicable.

Initialization

Create a mating scheme (do not use this base mating scheme, use one of its derived classes instead)

```
mating(numOffspring=1.0, numOffspringFunc=None, maxNumOffspring=0,
mode=MATE_NumOffspring, newSubPopSize=[], newSubPopSizeExpr="",
newSubPopSizeFunc=None)
```

By default, a mating scheme keeps a constant population size, generates one offspring per mating event. These can be changed using certain parameters. `newSubPopSize`, `newSubPopSizeExpr` and `newSubPopSizeFunc` can be used to specify subpopulation sizes of the offspring generation. `mode`, `numOffspring`, `maxNumOffspring` can be used to specify how many offspring will be produced at each mating event. This `mode` parameter can be one of

- `MATE_NumOffspring:` a fixed number of offspring at all mating events at this generation. If `numOffspring` is given, all generations use this fixed number. If `numOffspringFunc` is given, the number of offspring at each generation is determined by the value returned from this function.
- `MATE_NumOffspringEachFamily:` each family can have its own number of offspring. Usually, `numOffspringFunc` is used to determine the number of offspring of each family. If `numOffspring` is used, `MATE_NumOffspringEachFamily` is equivalent to `MATE_NumOffspring`.
- `MATE_GeometricDistribution:` a Geometric distribution with parameter `numOffspring` is used to determine the number of offspring of each family.
- `MATE_PoissonDistribution:` a Poisson distribution with parameter `numOffspring` is used to determine the number of offspring of each family.
- `MATE_BinomialDistribution:` a Binomial distribution with parameter `numOffspring` is used to determine the number of offspring of each family.
- `MATE_UniformDistribution:` a Uniform distribution `[a, b]` with parameter `numOffspring` (a) and `maxNumOffspring` (b) is used to determine the number of offspring of each family.

maxNumOffspring Used when `numOffspring` is generated from a binomial distribution

mode Can be one of `MATE_NumOffspring, MATE_NumOffspringEachFamily, MATE_GeometricDistribution, MATE_PoissonDistribution, MATE_BinomialDistribution, MATE_UniformDistribution`.

newSubPopSize An array of subpopulaitons sizes

newSubPopSizeExpr An expression that will return the new subpopulation size

newSubPopSizeFunc A function that accepts an `int` parameter(generation), an array of current population size and return an array of subpopulation sizes. This is usually easier to use than its expression version of this parameter.

numOffspring The number of offspring or *p* for a random distribution. Default to `1`. This parameter determines the number of offspring that a mating event will produce. Therefore, it determines the family size.

numOffspringFunc A Python function that returns the number of offspring or *p*

Member Functions

`x.clone()` Deep copy of a mating scheme

Example

Example 1.16: Demographic models and control of number of offspring per mating event

```
>>> # arbitrary demographic model
>>> def lin_inc(gen, oldsize=[]):
...     return [10+gen]*5
...
>>> simu = simulator(
...     population(subPop=[5]*5, loci=[1]),
...     randomMating(newSubPopSizeFunc=lin_inc)
... )
>>> simu.evolve(
...     ops = [
...         stat(popSize=True),
...         pyEval(r'"%d %d\n"%(gen, subPop[0]["popSize"])'),
...     ],
...     end=5
... )
0 10
1 11
2 12
3 13
4 14
5 15
True
>>>
>>> #
>>> # control the number of offspring per mating event
>>> # famSizes is only defined when DBG_MATING is defined
>>> TurnOnDebug(DBG_MATING)
>>> simu = simulator(population(50, loci=[1]),
...     randomMating(numOffspring=2,
...         maxNumOffspring=5,
...         mode=MATE_UniformDistribution))
>>> simu.step(ops=[])
True
>>> print simu.population(0).dvars().famSizes
[5, 5, 2, 5, 5, 5, 5, 3, 2, 5, 2, 3, 2, 1]
```

```
>>> TurnOffDebug(DBG_MATING)
Debug code DBG_MATING is turned off. cf. ListDebugCode(), TurnOnDebug().
>>>
```

1.8.2 Class noMating

A mating scheme that does nothing

Details

In this scheme, there is

- no mating. Parent generation will be considered as offspring generation.
- no subpopulation change. *During-mating* operators will be applied, but the return values are not checked. I.e., `subpopsizes` will be ignored although some during-mating operators may be applied.

Initialization

Creat a scheme with no mating

```
noMating(numOffspring=1.0, numOffspringFunc=None, maxNumOffspring=0,
mode=MATE_NumOffspring, newSubPopSize=[], newSubPopSizeExpr="",
newSubPopSizeFunc=None)
```

Note

All parameters are ignored!

Member Functions

`x.clone()` Deep copy of a scheme with no mating

1.8.3 Class binomialSelection

A mating scheme that uses binomial selection, regardless of sex

Details

No sex information is involved (binomial random selection). Offspring is chosen from parental generation by random or according to the fitness values. In this mating scheme,

- `numOffspring` protocol is honored;
- population size changes are allowed;
- selection is possible;
- haploid populaton is allowed.

Initialization

Create a binomial selection mating scheme

```
binomialSelection(numOffspring=1., numOffspringFunc=None,
maxNumOffspring=0, mode=MATE_NumOffspring, newSubPopSize=[],
newSubPopSizeExpr="", newSubPopSizeFunc=None)
```

Please refer to class `mating` for parameter descriptions.

Member Functions

`x.clone()` Deep copy of a binomial selection mating scheme

1.8.4 Class `randomMating`

A mating scheme of basic sexually random mating

Details

In this scheme, sex information is considered for each individual, and ploidy is always 2. Within each subpopulation, males and females are randomly chosen. Then randomly get one copy of chromosomes from father and mother. When only one sex exists in a subpopulation, a parameter (`contWhenUniSex`) can be set to determine the behavior. Default to continuing without warning.

Initialization

Create a random mating scheme

```
randomMating(numOffspring=1., numOffspringFunc=None,
maxNumOffspring=0, mode=MATE_NumOffspring, newSubPopSize=[],
newSubPopSizeFunc=None, newSubPopSizeExpr="", contWhenUniSex=True)
```

contWhenUniSex Continue when there is only one sex in the population, default to `true`

Please refer to class `mating` for descriptions of other parameters.

maxNumOffspring Used when `numOffspring` is generated from a binomial distribution

mode Can be one of `MATE_NumOffspring, MATE_NumOffspringEachFamily, MATE_GeometricDistribution, MATE_PoissonDistribution, MATE_BinomialDistribution`

newSubPopSize An array of subpopulation sizes, should have the same number of subpopulations as the current population

newSubPopSizeExpr An expression that will be evaluated as an array of subpopulation sizes

newSubPopSizeFunc An function that have parameter `gen` and `oldSize` (current subpopulation size)

numOffspring Number of offspring or *p* in some modes

numOffspringFunc A python function that determines the number of offspring or *p*

Member Functions

`x.clone()` Deep copy of a random mating scheme

1.8.5 Class `pyMating`

A Python mating scheme

Details

Hybird mating scheme. This mating scheme takes a Python function that accepts both the parental and offspring populations and this function is responsible for setting genotype, sex of the offspring generation. During-mating

operators, if needed, have to be applied from this function as well. Note that the subpopulaton size parameters are honored and the passed offspring generation has the desired (sub) population sizes. Parameters that control the number of offspring of each family are ignored.

This is likely an extremely slow mating scheme and should be used for experimental uses only. When a mating scheme is tested, it is recommended to implement it at the C++ level.

Initialization

Create a Python mating scheme

```
pyMating(func=None, newSubPopSize=[], newSubPopSizeExpr="",
newSubPopSizeFunc=None)
```

func A Python function that accepts two parameters: the parental and the offspring populations. The offspring population is empty, and this function is responsible for setting genotype, sex etc. of individuals in the offspring generation.

Member Functions

`x.clone()` Deep copy of a Python mating scheme

1.8.6 Determine the number of offspring during mating

The default value of `numOffspring` parameter makes a mating scheme produce one offspring per mating. This is the real random mating and should be used whenever possible. However, various situations require a larger family size or even changing the family size. simuPOP provides a comprehensive way to deal with this problem.

As described in the class reference, the method to determine the number of offspring is to set the `mode` parameter:

- `MATE_NumOffspring`: if `numOffspringFunc` is not given, the number of offspring will be the constant `numOffspring` all the time. Otherwise, `numOffspringFunc(gen)` will be called **once** for each generation to get the number of offspring for the matings happen in this generation.
- `MATE_NumOffspringEachFamily`: `numOffspringFunc` has to be given and will be called whenever a mating happens. Since `numOffspringFunc` can be **any** Python function, this mode allows arbitrary model of assigning the number of offspring during mating. The mode can be slow though.
- `MATE_GeometricDistribution`: `numOffspring` or the result of `numOffspringFunc` (evaluated at each generation) will be considered as p for a geometric distribution. The number of offspring for each mating is determined by
$$P(k) = p(1-p)^{k-1} \quad \text{for } k \geq 1$$
- `MATE_PoissonDistribution`: `numOffspring` or result of `numOffspringFunc` (evaluated at each generation) will be considered as p for a Poission distribution. The number of offspring for each mating is determined by
$$P(k) = \frac{p^{k-1}}{(k-1)!}e^{-p} \quad \text{for} k \geq 1$$
Since the mean of this shifted Poisson distribution is $p+1$, you need to specify, for example, 2, if you want a mean family size 3.
- `MATE_BinomialDistribution`: `numOffspring` or the result of `numOffspringFunc` (evaluated at each generation) will be considered as p for a Binomial distribution. Let N=`maxNumOffspring`, the number of offspring for each mating is determined by
$$P(k) = \frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k} \quad \text{for } N \geq k \geq 1$$

- `MATE_UniformDistribution: numOffspring` or the result of `numOffspringFunc` (evaluated at each generation), and `maxNumOffspring` will be considered as a, b for a Uniform distribution, respectively. The number of offspring for each mating is determined by

$$P(k) = \frac{1}{b-a} \text{ for } b \geq k \geq a$$

Note that all these distributions are adjusted to produce at least one offspring.

1.8.7 Determine subpopulation sizes of the next generation

The default behavior of simuPOP is to use the same population/subpopulation sizes as those of the parent generation. You can change this behavior by setting one of `newSubPopSize`, `newSubPopSizeExpr`, and `newSubPopSizeFunc` parameters:

- If you would like to have fixed subpopulation sizes, use `newSubPopSize=some_fixed_values`. This is useful when subpopulation sizes are changed by migration and you do want to keep constant subpopulation sizes.
- If subpopulation sizes can be easily calculated through an expression, you can use `newSubPopSizeExpr` to determine the new subpopulation sizes. For example, `newSubPopSizeExpr='[gen+10]'` uses the generation number + 10 as the new population size. More complicated expressions can be used, maybe along with `pyExec` operators, but in these cases, a specialized function and `newSubPopSizeFunc` are recommended.
- A more organized (and thus recommended) way to set new population/subpopulation sizes is through parameter `newSubPopSizeFunc`. To use this parameter, you need to define a Python function that takes two parameters: the generation number and the current subpopulation sizes, and return an array of new subpopulation sizes (return `[newsize]` instead of `newsize` when you do not have any subpopulation structure). The example of `class Mating` demonstrates the use of this parameter.

1.8.8 Demographic change functions

`newSubPopSizeFunc` can take a function with parameters `gen` and `oldSize`. A few functions are defined in `simuUtil.py` that will return such a function with given parameters. All these functions support a burnin stage and then split to equal sized subpopulations. For all these functions, you can test them by

```
func = oneOfTheDemographicFunc(parameters)
gen = range(0, yourEndGen)
r.plot(gen, [func(x)[0] for x in gen])
```

`numSubPop` is default to `1`. `split` is default to `0` or given `burnin` value. Population size change happens **after** burnin (start at `burnin+1`) and split happens at `split`.

```
ConstSize(size, split, numSubPop, bottleneckGen, bottleneckSize)
```

The population size is constant, but will split into `numSubPop` subpopulations at generation `split`. If `bottleneckGen` is specified, population size will be `bottleneckSize` at that generation.

```
LinearExpansion(initSize, endSize, end, burnin, split, numSubPop,
    bottleneckGen, bottleneckSize)
```

Linearly expand the population size from `intiSize` to `endSize` after `burnin`, split the population at generation `split`. If `bottleneckGen` is specified, population size will be `bottleneckSize` at that generation.

```
ExponentialExpansion(initSize, endSize, end, burnin, split,
    numSubPop, bottleneckGen, bottleneckSize)
```

Exponentially expand the population size from `intiSize` to `endSize` after `burnin`, split the population at generation `split`. If `bottleneckGen` is specified, population size will be `bottleneckSize` at that generation.

```
InstantExpansion(initSize, endSize, end, burnin, split,
    numSubPop, bottleneckGen, bottleneckSize)
```

Instaneously expand the population size from `intiSize` to `endSize` after `burnin`, split the population at generation `split`. If `bottleneckGen` is specified, population size will be `bottleneckSize` at that generation.

1.8.9 Sex chromosomes

Currently, only `randomMating()` in diploid population supports sex chromosomes. When `sexChrom()` is `False`, the sex of an offspring is determined randomly with probability $1/2$. Otherwise, it is determined by the existence of Y chromosome, I.e., what kind of sex chromosome an offspring get from his father.

Recombinations on sex chromosomes of females (XX) are just like those on autosomes. However, this is not true in males. Currently, recombinations between male sex chromosomes (XY) are *not* allowed (a bug/feature of recombinators). This may change later if exchanges of genes between pseudoautosomal regions of XY need to be modeled.

1.9 Operators

1.9.1 Class `baseOperator`

Base class of all classes that manipulate populations

Details

Operators are objects that act on populations. They can be applied to populations directly using their function forms, but they are usually managed and applied by a simulator.

There are three kinds of operators:

- built-in: written in C++, the fastest. They do not interact with Python shell except that some of them set variables that are accessible from Python.
- hybrid: written in C++ but calls a Python function during simulation. Less efficient. For example, a hybrid mutator pyMutator determines if an allele will be mutated and call a user-defined Python function to mutate it.
- pure Python: written in Python. The same speed as Python. For example, a `varPlotter` can plot Python variables that are set by other operators.

Operators can be applied at different stages of the life cycle of a generation. It is possible for an operator to apply multiple times in a life cycle. For example, a `savePopulation` operator might be applied before and after mating to trace parental information. More specifically, operators can be applied at *pre-*, *during-*, *post-mating*, or a combination of these stages. Applicable stages are usually set by default but you can change it by setting `stage=(PreMating|PostMating|DuringMating|PrePostMating)` parameter. Some operators ignore `stage` parameter because they only work at one stage.

Operators do not have to be applied at all generations. You can specify starting/ending generation (parameters `start`, `end`), gaps between applicable generations (parameter `step`), or specific generations (parameter `at`). For example, you might want to start applying migrations after certain burn-in generations, or calculate certain statistics only sparsely. Generation numbers can count from the last generation, using negative generation numbers.

Most operators are applied to every replicate of a simulator during evolution. However, you can apply operators to one (parameter `rep`) or a group of replicates only (parameter `grp`). For example, you can initialize different replicates

with different initial values and then start evolution. c.f. simulator::setGroup .Operators can have outputs, which can be standard (terminal) or a file. Output can vary with replicates and/or generations, and outputs from different operators can be accumulated to the same file to form table-like outputs.

Filenames can have the following format:

- `'filename'` this file will be overwritten each time. If two operators output to the same file, only the last one will succeed;
- `'>filename'` the same as `'filename'`;
- `'>>filename'` the file will be created at the beginning of evolution (`simulator::evolve`) and closed at the end. Outputs from several operators are appended;
- `'>>>filename'` the same as `'>>filename'` except that the file will not be cleared at the beginning of evolution if it is not empty;
- `'>'` standard output (terminal);
- `''` supress output.

The output filename does not have to be fixed. If parameter `outputExpr` is used (parameter `output` will be ignored), it will be evaluated when a filename is needed. This is useful when you need to write different files for different replicates/generations.

Initialization

Common interface for all operators (this base operator does nothing by itself.)

```
baseOperator(output, outputExpr, stage, begin, end, step, at, rep,
grp, infoFields)
```

at An array of active generations. If given, `stage`, `begin`, `end`, and `step` will be ignored.

begin The starting generation. Default to `0`. A negative number is allowed.

end Stop applying after this generation. A negative numbers is allowed.

grp Applicable group. Default to `GRP_ALL`. A group number for each replicate is set by `simulator.__init__` or `simulator::setGroup()`.

output A string of the output filename. Different operators will have different default `output` (most commonly `'>'` or `''`).

outputExpr An expression that determines the output filename dynamically. This expression will be evaluated against a population's local namespace each time when an output filename is required. For example, `"'>>out%s_%s.xml' % (gen, rep)"` will output to `>>>out1_1.xml` for replicate `1` at generation `1`.

rep Applicable replicates. It can be a valid replicate number, `REP_ALL` (all replicates, default), or `REP_LAST` (only the last replicate). `REP_LAST` is useful in adding newlines to a table output.

step The number of generations between active generations. Default to `1`.

Note

- Negative generation numbers are allowed for parameter `begin`, `end` and `at`. They are intepretted as `endGen + gen + 1`. For example, `begin = -2` in `simu.evolve(..., end=20)` starts at generation `19`.
- REP_ALL, REP_LAST, GRP_ALL are special constant that can only be used in the constructor of an operator. That is to say, explicit test of `rep() == REP_LAST` will not work.

Member Functions

`x.apply(pop)` Apply to one population. It does not check if the operator is activated.

`x.clone()` Deep copy of an operator

`x.diploidOnly()` Determine if the operator can be applied only for diploid population

`x.haploidOnly()` Determine if the operator can be applied only for haploid population

`x.infoField(idx)` Get the information field specified by user (or by default)

`x.infoSize()` Get the length of information fields for this operator

Example

Example 1.17: Common features of all operators

```
>>> simu = simulator(population(1),binomialSelection(), rep=2)
>>> op1 = pyOutput("a", begin=5, end=20, step=3)
>>> op2 = pyOutput("a", begin=-5, end=-1, step=2)
>>> op3 = pyOutput("a", at=[2,5,10])
>>> op4 = pyOutput("a", at=[-10,-5,-1])
>>> simu.evolve( [ pyEval(r"str(gen)+'\n'", begin=5, end=-1, step=2)],
...                                 end=10)
5
5
7
7
9
9
True
>>> #
>>> #
>>> # operator group
>>> from simuUtil import *
>>> simu = simulator(population(1),binomialSelection(), rep=4,
...     grp=[1,2,1,2])
>>> simu.step(
...     ops = [
...         pyEval(r"grp+3", grp=1),
...         pyEval(r"grp+6", grp=2),
...         pyOutput('\n', rep=REP_LAST)
...     ]
... )
4848
True
>>>
>>> #
>>> # parameter output
```

```
>>> simu = simulator(population(100), randomMating(), rep=2)
>>> simu.step(
...     preOps=[
...         initByFreq([0.2, 0.8], rep=0),
...         initByFreq([0.5, 0.5], rep=1) ],
...     ops = [
...         stat(alleleFreq=[0]),
...         pyEval('alleleFreq[0][0]', output='a.txt')
...     ]
... )
True
>>> # only from rep 1
>>> print open('a.txt').read()
0.455
>>>
>>> simu.step(
...     ops = [
...         stat(alleleFreq=[0]),
...         pyEval('alleleFreq[0][0]', output='>>a.txt')
...     ])
True
>>> # from both rep0 and rep1
>>> print open("a.txt").read()
0.230.46
>>>
>>> outfile='>>>a.txt'
>>> simu.step(
...     ops = [
...         stat(alleleFreq=[0]),
...         pyEval('alleleFreq[0][0]', output=outfile),
...         pyOutput("\t", output=outfile),
...         pyOutput("\n", output=outfile, rep=0)
...     ],
... )
True
>>> print open("a.txt").read()
0.230.460.27
0.415
>>> #
>>> # Output expression
>>> outfile="'>>a'+str(rep)+'.txt'"
>>> simu.step(
...     ops = [
...         stat(alleleFreq=[0]),
...         pyEval('alleleFreq[0][0]', outputExpr=outfile)
...     ]
... )
True
>>> print open("a0.txt").read()
0.255
>>> print open("a1.txt").read()
0.42
>>>
```

1.10 Simulator

1.10.1 Class `simulator`

Simulator manages several replicates of a population, evolve them using given mating scheme and operators

Details

Simulators combine three important components of simuPOP: population, mating scheme and operator together. A simulator is created with an instance of `population`, a replicate number `rep` and a mating scheme. It makes `rep` number of replicates of this population and control the evolutionary process of them.

The most important function of a simulator is `evolve()`. It accepts an array of operators as its parameters, among which, `preOps` and `postOps` will be applied to the populations at the beginning and the end of evolution, respectively, whereas `ops` will be applied at every generation.

A simulators separates operators into *pre-*, *during-*, and *post-mating* operators. During evolution, a simulator first apply all pre-mating operators and then call the `mate()` function of the given mating scheme, which will call during-mating operators during the birth of each offspring. After mating is completed, post-mating operators are applied to the offspring in the order at which they appear in the operator list.

Simulators can evolve a given number of generations (the `end` parameter of `evolve`), or evolve indefinitely until a certain type of operators called terminator terminates it. In this case, one or more terminators will check the status of evolution and determine if the simulation should be stopped. An obvious example of such a terminator is a fixation-checker.

A simulator can be saved to a file in the format of `'txt'`, `'bin'`, or `'xml'`. This allows youm to stop a simulator and resume it at another time or on another machine.

Initialization

Create a simulator

```
simulator(pop, matingScheme, stopIfOneRepStops=False,
applyOpToStoppedReps=False, rep=1, grp=[])
```

applyOpToStoppedReps If set, the simulator will continue to apply operators to all stopped replicates until all replicates are marked 'stopped'.

grp Group number for each replicate. Operators can be applied to a group of replicates using its `grp` parameter.

matingScheme A mating scheme

population A population created by `population()` function. This population will be copied `rep` times to the simulator. Its content will not be changed.

rep Number of replicates. Default to `1`.

stopIfOneRepStops If set, the simulator will stop evolution if one replicate stops.

Member Functions

`x.addInfoField(field, init=0)` Add an information field to all replicates

Add an information field to all replicate, and to the simulator itself. This is important because all populations must have the same genotypic information as the simulator. Adding an information field to one or more of the replicates will compromise the integrity of the simulator.

field information field to be added

x.addInfoFields(fields, init=0) Add information fields to all replicates

Add given information fields to all replicate, and to the simulator itself.

x.clone() Deep copy of a simulator

x.evolve(ops, preOps=[], postOps=[], end=-1, dryrun=False) Evolve all replicates of the population, subject to operators

Evolve to the `end` generation unless `end=-1`. An operator (terminator) may stop the evolution earlier.

`ops` will be applied to each replicate of the population in the order of:

- all pre-mating opertors
- during-mating operators called by the mating scheme at the birth of each offspring
- all post-mating operators If any pre- or post-mating operator fails to apply, that replicate will be stopped. The behavior of the simulator will be determined by flags `applyOpToStoppedReps` and `stopIfOneRepStopss`.

dryrun dryrun mode. Default to `False`.

end ending generation. Default to `-1`. In this case, there is no ending generation and a simulator will only be ended by a terminator. Otherwise, it should be a number greater than current generation number.

ops operators that will be applied at each generation, if they are active at that generation. (Determined by the `begin`, `end`, `step` and `at` parameters of the operator.)

postOps operators that will be applied after evolution. `evolve()` function will *not* check if they are active.

preOps operators that will be applied before evolution. `evolve()` function will *not* check if they are active.

Note: When `end = -1`, you can not specify negative generation parameters to operators. How would an operator know which genertion is the -1 genertion if no ending genertion is given?

x.gen() Return the current generation number

x.getPopulation(rep, destructive=False) Return a copy of population`rep`

By default return a cloned copy of population`rep` of the simulator. If `destructive==True`, the population is extracted from the simulator, leaving a defunct simulator.

destructive if true, destroy the copy of population within this simulator. Default to false. `getPopulation(rep, true)` is a more efficient way to get hold of a population when the simulator will no longer be used.

rep the index number of the replicate which will be obtained

x.group() Return group indexes

x.numRep() Return the number of replicates

x.population(rep) Return a reference to the `rep` replicate of this simulator.

rep the index number of replicate which will be accessed

Note: The returned reference is temporary in the sense that the refered population will be invalid after another round of evolution. If you would like to get a persistent population, please use `getPopulation(rep)`.

x.saveSimulator(filename, format="auto", compress=True) Save simulator in `'txt'`, `'bin'` or `'xml'` format

compress whether or not compress the file in `'gzip'` format

filename filename to save the simulator. Default to `simu`.

format format to save. Default to `auto`. I.e., determine the format by file extensions.

`x.setAncestralDepth(depth)` Set ancestral depth of all replicates

`x.setGen(gen)` Set the current generation. Usually used to reset a simulator.

`gen` new generation index number

`x.setGroup(grp)` Set groups for replicates

`x.setMatingScheme(matingScheme)` Set a new mating scheme

`x.step(ops=[], preOps=[], postOps=[], steps=1, dryrun=False)` Evolve `steps` generation

`x.vars(rep, subPop=-1)` Return the local namespace of population`rep`, equivalent to `x.population(rep).vars(subPop)`.

1.10.2 Generation number

Several aspects of the generation number may cause confusion:

- generation starts from zero
- a generation number presents a 'to-be-evolved' generation
- the ending generation specified in `evolve()` will be executed

That is to say, a new simulator will have generation 0 (at the beginning of generation 0). If you do `evolve(..., end=0), evolve` will evolve one generation and stop at the beginning of generation 1.

It may sound strange that

```
evolve(end=2)
```

evolve the population 3 generations. Generation 0, generation 1, and generation 2. When you use `start=0, step=5, end=10` for your operator, it will be applied at generations `0,5,10` etc. At the end of the simulation, current generation number is 3! (If you are familiar with C, this is like a `for` loop index). This is why you should test if a simulation is finished correctly by

```
if(simu.gen() == endGen+1)
```

instead of `simu.gen() == endGen.` (endGen is the value for parameter `end`).

1.10.3 Operator calling sequence

In a simulation, operators are applied at different stages, pre-, during-, and post-mating (controlled by `stage` parameter), at specified generations (controlled by `begin, end, step, at` parameters), and to specified replicates (controlled by `rep, grp` parameters). The order of applying operators usually does not matter but errors may occur if you are not careful. For example, `stat(...)` calculates the statistics of the current population. It is a pre-mating operator so you should set `stage=PostMating` and put it after all operators if you would like to measure a post-mating population. It also should be put before any operator (such as an terminator) that uses the shared variable set by `stat(...).`

If you are not sure about the calling sequence of operators, you can set the `dryrun` parameter of `evolve()` function to `True`. `evolve` will then print out the order of operators to apply. Consider that operators can be `PreMating, PostMating, PrePostMating, DuringMating` and the default value (parameter `stage`) may not be what you expect. Having a look at the calling sequence before the real evolution is always a good idea.

1.10.4 Save and Load

Using function `saveSimulator`, we can save a simulator to a file in the format of `'txt'`, `'bin'`, or `'xml'`. However, a mating scheme can not be saved and has to be re-specified in `LoadSimulator()`.

Example 1.18: save and load a simulator

```
>>> simu.saveSimulator("s.txt")
>>> simu.saveSimulator("s.xml", format="xml")
>>> simu.saveSimulator("s.bin", format="bin")
>>> simu1 = LoadSimulator("s.txt", randomMating())
>>> simu2 = LoadSimulator("s.xml", randomMating(), format="xml")
>>> simu3 = LoadSimulator("s.bin", randomMating(), format="bin")
>>>
```

1.11 Population variables

Populations are associated with Python variables. These variables are usually set by various operators but you can also set them manually. For example, `stat` operator calculates many population statistics and store the results in a population's local namespace.

1.11.1 `vars()`and `dvars()` functions

Conceptually, population variables are organized as follows (looking from a simulator's point of view):

```
simu.vars(0)        simu.vars(1) ...     // replicate
  popSize             popSize            // local namespace
  alleleFreq[0]       alleleFreq[0]      // allele frequency at locus 1
  alleleFreq[1]       alleleFreq[1]      // at locus 2
  ...                 ....
  subPop[0]           subPop[0]          // subpop namespace
    popSize             popSize          // subpopulation 1 size
    allaleFreq[0]       allaleFreq[0]    // allele frequency at locus 1
    ...                 ...
  subPop[1]           subPop[1]          // variables for subpop 2
    ...                 ...
```

You can refer to these variables using `population::vars()` or `population::dvars()` function. The returned values of `vars()` and `dvars()` reflect the same dictionary, but `dvars()` uses a little Python magic so that you can use attribute syntax to access dictionary keys. Because `a.allaleFreq[0]` is easier to read than `a['alleleFreq'][0]`, `dvars()` is more frequently used.

There are several ways to use these two functions

- `pop.vars(), pop.dvars()` return the variables of population `pop`
- `pop.vars(subPop), pop.dvars(subPop)` returns dictionary `pop.vars()['subPop'][subPop]`
- `simu.vars(rep), simu.dvars(rep)` return the variables of the `rep`'th population of simulator `simu` , i.e. `simu.population(rep).vars()`.
- `simu.vars(rep, subPop), simu.dvars(rep, subPop)` returns dictionary `simu.vars(rep)['subPop'][subPop]`

Direct access to variables `pop.vars()['subPop'][subPop]` is provided because statistics calculator `stat`, by default, calculates the same set of statistics for all subpopulations (and the whole population).

To have a look at all variables defined in this dictionary, you can use function `ListVars` defined in `simuUtil.py`. With wxPython installed, this function opens a nice window with a tree representing the variables. Without wxPython (or use parameter `useWxPython=False`), variables are displayed in an indented form. Several parameters can be used to limit your display. They are

- `level`: the level of the tree, further nested variables will not be displayed
- `name`: the name of the variable to display
- `subPop`: whether or not display variables for each subpopulation.

Example 1.19: Population variables

```
>>> from simuUtil import ListVars
>>> pop = population(subPop=[1000, 2000], loci=[1])
>>> InitByFreq(pop, [0.2, 0.8])
>>> ListVars(pop.vars(), useWxPython=False)
 rep :  -1
 grp :  -1
>>> Stat(pop, popSize=1, alleleFreq=[0])
>>> # subPop is True by default, use name to limit the variables to display
>>> ListVars(pop.vars(), useWxPython=False, subPop=False, name='alleleFreq')
 alleleFreq :
   [0]
     [0]        0.199833333333
     [1]        0.800166666667
>>> # print number of allele 1 at locus 0
>>> print pop.vars()['alleleNum'][0][1]
4801
>>> print pop.dvars().alleleNum[0][1]
4801
>>> print pop.dvars().alleleFreq[0]
[0.19983333333333334, 0.80016666666666669]
>>> print pop.dvars(1).alleleNum[0][1]
3196
>>>
```

1.11.2 Local namespace, `pyEval` and `pyExec` operators

Population variables is a Python dictionary, and furthermore a *Local namespace*, which means that you can use dictionary items as variables during evaluation. To evaluate in a population's local namespace, you can use function `population::evalulate()` or `population::execute()`. For example:

Example 1.20: Local namespaces of populations

```
>>> pop = population(subPop=[1000, 2000], loci=[1])
>>> InitByFreq(pop, [0.2, 0.8])
>>> Stat(pop, popSize=1, alleleFreq=[0])
>>> print pop.evaluate('alleleNum[0][0] + alleleNum[0][1]')
6000
>>> pop.execute('newPopSize=int(popSize*1.5)')
>>> ListVars(pop.vars(), level=1, useWxPython=False)
```

```
 newPopSize :   4500
 grp :  -1
 rep :  -1
 popSize :      3000
 numSubPop :    2
 alleleNum :
   list of length 1
 subPopSize :
   list of length 2
 alleleFreq :
   list of length 1
 subPop
   list of length 2
>>> # this variable is 'local' to the population and is
>>> # not available in the main namespace
>>> newPopSize
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
NameError: name 'newPopSize' is not defined
>>> #
>>> simu = simulator(population(10),noMating(), rep=2)
>>> # evaluate an expression in different areas
>>> print simu.vars(1)
{'rep': 1, 'gen': 0, 'grp': 1}
>>> print simu.population(0).evaluate("grp*2")
0
>>> print simu.population(1).evaluate("grp*2")
2
>>> # a statement (no return value)
>>> simu.population(0).execute("myRep=2+rep*rep")
>>> simu.population(1).execute("myRep=2*rep")
>>> print simu.vars(0)
{'rep': 0, 'myRep': 2, 'gen': 0, 'grp': 0}
>>>
```

These two functions are rarely used, because

```
pop.evaluate('alleleNum[0][1] + 1')
```

is equivalent to

```
pop.dvar().alleleNum[0][1] + 1
```

Operators `pyEval/pyExec` are more useful in that they can be applied to different populations during evolution, and report statistics calculated by operator `stat` dynamically. The difference between these two operators are that `pyEval` evaluates a Python expression and returns its value, while `pyExec` executes a list of statements in the form of a multi-line string, and does not return any value.

Example 1.21: Use of operators pyEval and pyExec

```
>>> simu = simulator(population(100, loci=[1]),
...     randomMating(), rep=2)
>>> simu.evolve(
...     preOps = [initByFreq([0.2, 0.8])],
...     ops = [ stat(alleleFreq=[0]),
```

```
...         pyExec('myNum = alleleNum[0][0] * 2'),
...         pyEval(r'"gen %d, rep %d, num %d, myNum %d\n"' \
...             ' % (gen, rep, alleleNum[0][0], myNum)')
...         ],
...     end=2
... )
gen 0, rep 0, num 43, myNum 86
gen 0, rep 1, num 30, myNum 60
gen 1, rep 0, num 35, myNum 70
gen 1, rep 1, num 40, myNum 80
gen 2, rep 0, num 41, myNum 82
gen 2, rep 1, num 41, myNum 82
True
>>>
```

1.12 Information fields

An individuals have genotype, sex and affection status information, but other information may be needed. For example, one or more trait values may be needed to calculate quantitative traits, and one may want to keep track of all offspring of a parent. Because the need for information fields varies from simulation to simulation, simuPOP does not fix the amount of information fields, and allow users to specify these fields during the construction of populations, or add them when you need them.

Operators may require certain information fields to work properly. For example, all selectors require field `fitness` to store evaluated fitness values for each individual. `parentTagger` needs `father_idx` and `mother_idx` to store indices of the parents of each individual in the parental generation. These information fields can be added by the `infoFields` parameter of the population constructor or be added later using relevant function. If a required information field is unavailable, an error message will appear and tell you which field is needed. Some operators allow you to specify which information field(s) to use. For example, quantitative trait operator can work on specified fields so an individual can have several quantitative traits.

The information fields is usually set during population creation, using the `infoFields` option of population constructor. It can also be set or added by functions

- `pop.setInfoFields(fields, init)` set information fields of a population, removing all previous ones
- `pop.addInfoField(field, init)` add an information field to a population
- `pop.addInfoFields(fields, init)` add information fields to a population
- `simu.addInfoField(field, init)` add an information field to all populations in a simulator
- `simu.addInfoFields(fields, init)` add information fields to all populations in a simulator

When adding information fields to a simulator, information fields are added to all populations of the simulator. Note that it is illegal to add information field (or in a broader sense changing genotypic structure) to part of the populations of a simulator, because all populations in a simulator should have the same genotypic structure.

One can read/write information fields at individual level:

- `ind.info(idx), ind.info(name)` return individual information field by index or name
- `ind.setInfo(value, idx), ind.setInfo(value, name)` set individual information field by index or name

- `ind.arrInfo()` returns a carray of all information fields of an individual

or at the population level

- `pop.indInfo(idx, order), pop.indInfo(name, order)` return an information field (referred by index or name) of all individuals
- `pop.indInfo(idx, subPop, order), pop.indInfo(name, subPop, order)` return an information field (referred by index or name) of all individuals in a subpopulation `subPop`.
- `pop.setIndInfo(values, idx, order), pop.setInfIndo(values, name, order)` set information fields of all individuals with values in an array.
- `pop.arrIndInfo(order)` return an carray of all information fields
- `pop.arrIndInfo(subPop, order)` return an carray of all information fields of subpopulation `subPop`.

Both `idx` or `name` can be used in these functions. `name` is easier to use but `idx`, which can be obtained by `idx=pop.infoIdx(name),` is faster. Although population information fields are kept in a population object linearly, there is no guarantee that they are ordered. If you would like to access `info` individual by individual, passing `order=True` to these functions will ensure that the returned information fields are ordered by individual order. If you only need to get a summary of some information fields, passing `order=False` will speed up the process.

`ind.arrInfo()` returns an carray `f1, f2, f3` (assuming `infoSize()==3`) of individual `ind`. At a population level, `pop.arrIndInfo([subPop], order)` returns a carray of `f1, f2, f3, f1, f2, f3, ...` of individual information fields, which are not necessarily in the order of individuals unless `order = True` is set. `indInfo` is more convenient but it is less efficient than `arrIndInfo`. For example, the following two examples both assign an information field from the value of another one, but **??** is more efficient.

Example 1.22: Use regular information field function

```
>>> pop = population(10, infoFields=['a', 'b'])
>>> aIdx = pop.infoIdx('a')
>>> bIdx = pop.infoIdx('b')
>>> for ind in pop.individuals():
...     a = ind.info(aIdx)
...     ind.setInfo(a+1, bIdx)
...
>>> print pop.indInfo(bIdx, False)
(1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)
>>>
```

Example 1.23: Use carray information fields function

```
>>> pop = population(10, infoFields=['a', 'b'])
>>> aIdx = pop.infoIdx('a')
>>> bIdx = pop.infoIdx('b')
>>> info = pop.arrIndInfo(False)
>>> sz = pop.infoSize()
>>> for idx in range(pop.popSize()):
...     info[sz*idx + bIdx] = info[sz*idx + aIdx] + 1
...
>>> print pop.indInfo(bIdx, False)
(1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)
>>>
```

CHAPTER

TWO

Operator References

This chapter will list all functions, types and operators by category. The reference for `class baseOperator` is in section **??**.

2.1 Python operators

A Python operator that works directly on simuPOP population or individuals.

2.1.1 Class `pyOperator`

A python operator that directly operate a population.

Details

This operator accepts a function that can take the form of

- `func(pop)` when `stage=PreMating` or `PostMating`, without setting `param`;
- `func(pop, param)` when `stage=PreMating` or `PostMating`, with `param`;
- `func(pop, off, dad, mom)` when `stage=DuringMating` and `passOffspringOnly=False`, without setting `param`;
- `func(off)` when `stage=DuringMating` and `passOffspringOnly=True`, and without setting `param`;
- `func(pop, off, dad, mom, param)` when `stage=DuringMating` and `passOffspringOnly=False`, with `param`;
- `func(off, param)` when `stage=DuringMating` and `passOffspringOnly=True`, with `param`.

For `Pre-` and `PostMating` usages, a population and an optional parameter is passed to the given function. For `DuringMating` usages, population, offspring, its parents and an optional parameter are passed to the given function. Arbitrary operations can be applied to the population and offspring (if `stage=DuringMating`).

Initialization

Python operator, using a function that accepts a population object.

```
pyOperator(func, param=None, stage=PostMating, formOffGenotype=False,
passOffspringOnly=False, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

formOffGenotype This option tells the mating scheme this operator will set the genotype of offspring (valid only for `stage=DuringMating`). By default (`formOffGenotype=False`), a mating scheme will set the genotype of offspring before it is passed to the given Python function. Otherwise, a 'blank' offspring will be passed.

func A Python function. Its form is determined by other parameters.

param Any Python object that will be passed to `func` after `pop` parameter. Multiple parameters can be passed as a tuple.

passOffspringOnly If `True`, `pyOperator` will expect a function of form `func(off [,param])`, instead of `func(pop, off, dad, mom [, param])` which is used when `passOffspringOnly` is `False`. Because many during-mating `pyOperator` only need access to offspring, this will improve efficiency. Default to `False`.

Note

- Output to `output` or `outputExpr` is not supported. That is to say, you have to open/close/append to files explicitly in the Python function.
- This operator can be applied `Pre-`, `During-` or `Post-` mating and is applied `PostMating` by default. For example, if you would like to examine the fitness values set by a selector, a `PreMating` Python operator should be used.

Member Functions

`x.apply(pop)` Apply the `pyOperator` operator to one population

`x.clone()` Deep copy of a `pyOperator` operator

A Python operator accepts a function and an optional parameter. When `pyOperator` is called, it will simply pass the accepted population (or parents and offspring in the case of `stage=DuringMating`) to the function. To use this operator, in case of `stage=PostMating`, you will need to

- define a function that handle a population as you wish.

```
def myOperator(pop, para):
    'do whatever you want'
    return True
```

  If you return `False`, this operator will work like a terminator.

- use `pyOperator` in the form of

```
pyOperator(mfunc=pyOperator, param=para)
```

  all parameters of an operator are supported except for `output` and `outputExpr` which are ignored for now.

This operator allows implementation of arbitrarily complicated operators,. To use this operator, you will have to know how to use population-related functions. The following example shows how to implement a dynamic mutator which mutate loci according to their allele frequencies.

Example 2.1: define a python operator

```
>>> def dynaMutator(pop, param):
...     ''' this mutator mutate common loci with low mutation rate
...     and rare loci with high mutation rate, as an attempt to
```

```
...     bring allele frequency of these loci at an equal level.'''
...     # unpack parameter
...     (cutoff, mu1, mu2) = param;
...     Stat(pop, alleleFreq=range( pop.totNumLoci() ) )
...     for i in range( pop.totNumLoci() ):
...         # 1-freq of wild type = total disease allele frequency
...         if 1-pop.dvars().alleleFreq[i][1] < cutoff:
...             KamMutate(pop, maxAllele=2, rate=mu1, loci=[i])
...         else:
...             KamMutate(pop, maxAllele=2, rate=mu2, loci=[i])
...     return True
... #end
...
```

Example 2.2: use of python operator

```
>>> pop = population(size=10000, ploidy=2, loci=[2, 3])
>>>
>>> simu = simulator(pop, randomMating())
>>>
>>> simu.evolve(
...     preOps = [
...         initByFreq( [.6, .4], loci=[0,2,4]),
...         initByFreq( [.8, .2], loci=[1,3]) ],
...     ops = [
...         pyOperator( func=dynaMutator, param=(.5, .1, 0) ),
...         stat(alleleFreq=range(5)),
...         pyEval(r'"%f\t%f\n"%(alleleFreq[0][1],alleleFreq[1][1])', step=10)
...         ],
...     end = 30
... )
0.398800        0.198200
0.396850        0.199450
0.399500        0.204500
0.387550        0.203450
True
>>>
```

`pyOperator` can also be a during-mating operator. You will need to define a function

```
def Func(pop, off, dad, mom, para)
```

or

```
def shortFunc(off, para)
```

where `para` can be ignored. To use this operator, you can do

```
pyOperator(stage=DuringMating, func=Func, param=someparam, formOffGenotype=True)
```

or

```
pyOperator(stage=DuringMating, func=shortFunc, param=someparam,
formOffGenotype=False, passOffspringOnly=True)
```

If your during-mating `pyOpeartor` returns `False`, the individual will be discarded. Therefore, you can write a filter in this way. However, since the Python function will be called for each mating event, the cost of using such an operator is high, especially when population size is large.

An example of during-mating `pyOperator` can be found in `scripts/demoPyOperator.py`.

2.1.2 Class `pyIndOperator`

Individual operator

Details

This operator is similar to a `pyOperator` but works at the individual level. It expects a function that accepts an individual, optional genotype at certain loci, and an optional parameter. When it is applied, it passes each individual to this function. When `infoFields` is given, this function should return an array to fill these infoFields. Otherwise, True/False is expected.More specifically, `func` can be

- `func(ind)` when neither loci nor param is given.
- `func(ind, genotype)` when loci is given
- `func(ind, param)` when param is given
- `func(ind, genotype, param)` when both loci and param are given.

Initialization

A `Pre-` or `PostMating` Python operator that apply a function to each individual

```
pyIndOperator(func, loci=[], param=None, stage=PostMating,
formOffGenotype=False, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

func A Python function that accepts an individual and optional genotype and parameters.

infoFields If given, `func` is expected to return an array of the same length and fill these `infoFields` of an individual.

param Any Python object that will be passed to `func` after `pop` parameter. Multiple parameters can be passed as a tuple.

Member Functions

`x.apply(pop)` Apply the `pyIndOperator` operator to one population

`x.clone()` Deep copy of a `pyIndOperator` operator

2.2 Initialization

2.2.1 Class `initializer`

Initialize alleles at the start of a generation

Details

Initializers are used to initialize populations before evolution. They are set to be `PreMating` operators by default. simuPOP provides three initializers. One assigns alleles by random, one assigns a fixed set of genotypes, and the last one calls a user-defined function.

Initialization

Create an initializer. default to be always active

```
initializer(subPop=[], indRange=[], loci=[], atPloidy=-1,
maleFreq=0.5, sex=[], stage=PreMating, begin=0, end=-1, step=1,
at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

atPloidy Initialize which copy of chromosomes. Default to all.

indRange A `[begin, end]` pair of the range of absolute indexes of individuals, for example, `([1,2])`; or an array of `[begin, end]` pairs, such as `([[1,4],[5,6]])`. This is how you can initialize individuals differently within subpopulations. Note that ranges are in the form of [a,b). I.e., range [4,6] will intialize individual 4, 5, but not 6. As a shortcut for [4,5], you can use [4] to specify one individual.

loci A vector of locus indexes at which initialization will be done. If empty, apply to all loci.

locus A shortcut to `loci`

maleFreq Male frequency. Default to `0.5`. Sex will be initialized with this parameter.

subPop An array specifies applicable subpopulations

Member Functions

`x.clone()` Deep copy of an initializer

2.2.2 Class `initByFreq` (Function form: `InitByFreq`)

Initialize genotypes by given allele frequencies, and sex by male frequency

Details

This operator assigns alleles at `loci` with given allele frequencies. By default, all individuals will be assigned with random alleles. If `identicalInds=True`, an individual is assigned with random alleles and is then copied to all others. If `subPop` or `indRange` is given, multiple arrays of alleleFreq can be given to given different frequencies for different subpopulation or individual ranges.

Initialization

Randomly assign alleles according to given allele frequencies

```
initByFreq(alleleFreq=[], identicalInds=False, subPop=[],
indRange=[], loci=[], atPloidy=-1, maleFreq=0.5, sex=[],
stage=PreMating, begin=0, end=1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

alleleFreq An array of allele frequencies. The sum of all the frequencies must be 1; or for a matrix of allele frequencies, each row corresponses to a subpopulation or range.

identicalInds Whether or not make individual genotypies identical in all subpopulation. If `True`, this operator will randomly generate genotype for an individual and spread it to the whole subpopulation in the given range.

sex An array of sex `[Male, Female, Male...]` for individuals. The length of sex will not be checked. If it is shorter than the number of individuals, sex will be reused from the beginning.

stage Default to `PreMating`

Member Functions

x.apply(pop) Apply this operator to population`pop`

x.clone() Deep copy of the operator `initByFreq`

Example

Example 2.3: Operator initByFreq

```
>>> simu = simulator(
...     population(subPop=[2,3], loci=[5,7], maxAllele=1),
...     randomMating(), rep=1)
>>> simu.step([
...     initByFreq(alleleFreq=[ [.2,.8],[.8,.2]]),
...     dumper(alleleOnly=True)
...   ])
individual info:
sub population 0:
   0: MU 00111 0111110 | 10111 0111101
   1: MU 11111 1110111 | 10111 1011111
sub population 1:
   2: MU 00000 0001000 | 00110 0000000
   3: FU 00000 0001000 | 01000 0000000
   4: MU 00000 0001000 | 01000 0000000
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.2.3 Class `initByValue` (Function form: `InitByValue`)

Initialize genotype by value and then copy to all individuals

Details

`INITBYVALUE` operator gets one copy of chromosomes or the whole genotype (or of those corresponds to `loci`) of an individual and copy them to all or a subset of individuals. This operator assign given alleles to specified individuals. Every individual will have the same genotype. The parameter combinations should be

- `value - subPop/indRange`: individual in `subPop` or in range(s) will be assigned genotype 'value';
- `subPop/indRange`: `subPop` or `indRange` should have the same length as values. Each item of values will be assigned to each `subPop` or `indRange`.

Initialization

Initialize populations by given alleles

```
initByValue(value=[], loci=[], atPloidy=-1, subPop=[], indRange=[],
proportions=[], maleFreq=0.5, sex=[], stage=PreMating, begin=0,
end=1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

maleFreq Male frequency

proportions An array of percentages for each item in values. If given, assign given genotypes randomly.

sex An array of sex `[Male, Female, Male...]` for individuals. The length of sex will not be checked. If length of sex is shorter than the number of individuals, sex will be reused from the beginning.

stages Default to `PreMating`

value An array of genotypes of one individual, having the same length as the length of `loci()` or `loci()*ploidy()` or `pop.genoSize()` (whole genotype) or `totNumLoci()` (one copy of chromosome). This parameter can also be an array of arrays of genotypes of one individual. Should have length one or equal to `subpop` or ranges or proportion. If value is an array of values, it should have the same length as `subpop`, `indRange` or `proportions`.

Member Functions

`x.apply(pop)` Apply this operator to population`pop`

`x.clone()` Deep copy of the operator `initByValue`

Example

Example 2.4: Operator initByValue

```
>>> simu = simulator(
...     population(subPop=[2,3], loci=[5,7], maxAllele=9),
...     randomMating(), rep=1)
>>> simu.step([
...     initByValue([1]*5 + [2]*7 + [3]*5 +[4]*7),
...     dumper(alleleOnly=True)])
individual info:
sub population 0:
   0: MU 11111 2222222 | 33333 4444444
   1: FU 33333 4444444 | 11111 2222222
sub population 1:
   2: FU 33333 4444444 | 11111 2222222
   3: FU 11111 2222222 | 33333 2222222
   4: MU 33333 2222222 | 11111 2222222
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.2.4 Class `spread` (Function form: `Spread`)

Copy the genotype of an individual to all individuals

Details

`SPREAD(IND, SUBPOP)` spreads the genotype of `ind` to all individuals in an array of subpopulations. The default value of `subPop` is the subpopulation where `ind` resides.

Initialization

Copy genotypes of `ind` to all individuals in `subPop`

```
spread(ind, subPop=[], stage=PreMating, begin=0, end=1, step=1,
at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply this operator to population`pop`

`x.clone()` Deep copy of the operator `spread`

2.2.5 Class `pyInit` (Function form: `PyInit`)

A python operator that uses a user-defined function to initialize individuals.

Details

`PYINIT` is a hybrid initializer. User should define a function with parameters allele, ploidy and subpopulation indexes, and return an allele value. Users of this operator must supply a Python function with parameter `(index, ploidy, subpop)`. This operator will loop through all individual in each subpopulation and call this function to initialize populations. The arrange of parameters allows different initialization scheme for each subpop.

Initialization

Initialize populations using given user function

```
pyInit(func, subPop=[], loci=[], atPloidy=-1, indRange=[],
maleFreq=0.5, sex=[], stage=PreMating, begin=0, end=1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

atPloidy Initialize which copy of chromosomes. Default to all.

func A Python function with parameter `(index, ploidy, subpop)`, where

- `index` is the allele index ranging from `0` to `totNumLoci(-1)`,
- `ploidy` is the index of the copy of chromosomes)
- `subpop` is the subpopulation index.

The return value of this function should be an integer.

loci A vector of loci indexes. If empty, apply to all loci.

locus A shortcut to `loci`

stage Default to `PreMating`

Member Functions

`x.apply(pop)` Apply this operator to population`pop`

`x.clone()` Deep copy of the operator `pyInit`

Example

Example 2.5: Operator pyInit

```
>>> def initAllele(ind, p, sp):
...   return sp + ind + p
...
>>> simu = simulator(
...     population(subPop=[2,3], loci=[5,7]),
...     randomMating(), rep=1)
```

```
>>> simu.step([
...     pyInit(func=initAllele),
...     dumper(alleleOnly=True, dispWidth=2)])
individual info:
sub population 0:
   0: FU   0  1  2  3  4   5  6  7  8  9 10 11 |   0  1  2  3  4   6  7
8  9 10 11 12
   1: FU   1  2  3  4  5   5  6  7  8  9 10 11 |   0  1  2  3  4   6  7
8  9 10 11 12
sub population 1:
   2: MU   1  2  3  4  5   6  7  8  9 10 11 12 |   1  2  3  4  5   7  8
9 10 11 12 13
   3: MU   1  2  3  4  5   6  7  8  9 10 11 12 |   2  3  4  5  6   6  7
8  9 10 11 12
   4: MU   2  3  4  5  6   6  7  8  9 10 11 12 |   1  2  3  4  5   6  7
8  9 10 11 12
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.3 Migration

2.3.1 Class `migrator`

Migrate individuals from a (sub) population to another (sub) population

Details

Migrator is the only way to mix genotypes of several subpopulations because mating is strictly within subpopulations in simuPOP. Migrators are quite flexible in simuPOP in the sense that

- Migration can happen from and to a subset of subpopulations.
- Migration can be done by probability, proportion or by counts. In the case of probability, if the migration rate from subpopulation `a` to `b` is `r`, then everyone in subpopulation `a` will have this probability to migrate to `b`. In the case of proportion, exactly `r*size_of_subPop_a` individuals (chosen by random) will migrate to subpopulation `b`. In the last case, a given number of individuals will migrate.
- New subpopulation can be generated through migration. You simply need to migrate to a new subpopulation number.

Initialization

Create a migrator

```
migrator(rate, mode=MigrByProbability, fromSubPop=[], toSubPop=[],
stage=PreMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

fromSubPop An array of 'from' subpopulations. Default to all. If a single `subpop` is specified, `[]` can be ignored. I.e., `[a]` is equvalent to `a`.

mode One of `MigrByProbability` (default), `MigrByProportion` or `MigrByCounts`

rate Migration rate, can be a proportion or counted number. Determined by parameter `mode`. `rate` should be an m by n matrix. If a number is given, the migration rate will be a `m` by `n` matrix of value `r`

stage Default to `PreMating`

toSubPop An array of 'to' subpopulations. Default to all subpopulations. If a single `subpop` is specified, `[]` can be ignored.

Note

- The overall population size will not be changed. (Mating schemes can do that). If you would like to keep the subpopulation size after migration, you can use the `newSubPopSize` or `newSubPopSizeExpr` parameter of a mating scheme.
- `rate` is a matrix with dimensions determined by `fromSubPop` and `toSubPop`. By default, `rate` is a matrix with element `r(i,j)`, where `r(i, j)` is the migration rate, probability or count from subpopulation `i` to `j`. If `fromSubPop` and/or `toSubPop` are given, migration will only happen between these subpopulations. An extreme case is 'point migration', `rate=[[r]], fromSubPop=a, toSubPop=b` which migrate from subpopulation `a` to `b` with given rate `r`.

Member Functions

`x.apply(pop)` Apply the `migrator`

`x.clone()` Deep copy of a migrator

`x.rate()` Return migration rate

`x.setRates(rate, mode)` Set migration rate

Format should be `0-0 0-1 0-2, 1-0 1-1 1-2, 2-0, 2-1, 2-2`. For mode `MigrByProbability` or `MigrByProportion`, `0-0,1-1,2-2` will be set automatically regardless of input.

2.3.2 Functions (Python) `MigrIslandRates, MigrStepstoneRates` (simuUtil.py)

Migrator is very flexible. It can accept arbitrary migration matrix, from any subset of subpopulations to any (even new) other subset of subpopulations. To facilitate the use of common theoretical migration models, several functions are defined in `simuUtil.py`.

- `MigrIslandRates(r, n)` returns a migration matrix

$$\begin{pmatrix} 1-r & \frac{r}{n-1} & \dots & \dots & \frac{r}{n-1} \\ \frac{r}{n-1} & 1-r & \dots & \dots & \frac{r}{n-1} \\ & & \dots & & \\ \frac{r}{n-1} & \dots & \dots & 1-r & \frac{r}{n-1} \\ \frac{r}{n-1} & \dots & \dots & \frac{r}{n-1} & 1-r \end{pmatrix}$$

- `MigrStepstoneRates(r, n, circular=False)` returns a migration matrix

$$\begin{pmatrix} 1-r & r & & & \\ r/2 & 1-r & r/2 & & \\ & & \dots & & \\ & & r/2 & 1-r & r/2 \\ & & & r & 1-r \end{pmatrix}$$

and if `circular=True`, returns

$$\begin{pmatrix} 1-r & r/2 & & & r/2 \\ r/2 & 1-r & r/2 & & \\ & & \ldots & & \\ & & r/2 & 1-r & r/2 \\ r/2 & & & r/2 & 1-r \end{pmatrix}$$

2.3.3 Class `pyMigrator`

A more flexible Python migrator

Details

This migrator can be used in two ways

- define a function that accepts a generation number and returns a migration rate matrix. This can be used in the varying migration rate cases.
- define a function that accepts individuals etc, and returns the new subpopulation ID.

More specifically, `func` can be

- `func(ind)` when neither `loci` nor `param` is given.
- `func(ind, genotype)` when `loci` is given.
- `func(ind, param)` when `param` is given.
- `func(ind, genotype, param)` when both `loci` and `param` are given.

Initialization

Create a hybrid migrator

```
pyMigrator(rateFunc=None, mode=MigrByProbability, fromSubPop=[],
toSubPop=[], indFunc=None, loci=[], param=None, stage=PreMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

indFunc A Python function that accepts an individual, optional genotype and parameter, then returns a subpopulation id. This method can be used to separate a population according to individual genotype.

rateFunc A Python function that accepts a generation number, current subpopulation sizes, and returns a migration rate matrix. The migrator then migrate like a usual migrator.

stage Default to `PreMating`

Member Functions

`x.apply(pop)` Apply a `pyMigrator`

`x.clone()` Deep copy of a `pyMigrator`

2.3.4 Class `splitSubPop` (Function form: `SplitSubPop`)

Split a subpopulation

Initialization

Split a subpopulation or the whole population as subpopulation `0`

```
splitSubPop(which=0, sizes=[], proportions=[], subPopID=[],
randomize=True, stage=PreMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

proportions Proportions of new subpopulations. Should be added up to `1`.

sizes New subpopulation sizes. The sizes should be added up to the original subpopulation (subpopulation `which`) size.

subPopID New subpopulation IDs. Otherwise, the operator will automatically set new subpopulation IDs to new subpopulations.

which Which subpopulation to split. If there is no subpopulation structure, use 0 as the first (and only) subpopulation.

Member Functions

`x.apply(pop)` Apply a `splitSubPop` operator

`x.clone()` Deep copy of a `splitSubPop` operator

2.3.5 Class `mergeSubPops` (Function form: `MergeSubPops`)

Merge subpopulations

Details

This operator merges subpopulations `subPops` to a single subpopulation. If `subPops` is ignored, all subpopulations will be merged.

Initialization

Merge subpopulations

```
mergeSubPops(subPops=[], removeEmptySubPops=False, stage=PreMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

subPops Subpopulatinos to be merged. Default to all.

Member Functions

`x.apply(pop)` Apply a `mergeSubPops` operator

`x.clone()` Deep copy of a `mergeSubPops` operator

2.4 Mutation

2.4.1 Class `mutator`

Base class of all mutators.

Details

The base class of all functional mutators. It is not supposed to be called directly.Every mutator can specify `rate` (equal rate or different rates for different loci) and a vector of applicable loci (default to all but should have the same length as `rate` if `rate` has length greater than one).Maximum allele can be specified as well but more parameter, if needed, should be implemented by individual mutator classes.There are number of possible allelic states. Most theoretical studies assume an infinite number of allelic states to avoid any homoplasy. If it facilitates any analysis, this is however extremely unrealistic.

Initialization

Create a mutator

```
mutator(rate=[], loci=[], maxAllele=0, output=">", outputExpr="",
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

All mutators have the following common parameters. However, the actual meaning of these parameters may vary according to different model. The only differences between the following mutators are they way they actually mutate an allele, and corresponding input parameters. The number of mutation events at each locus is recorded and can be accessed from the `mutationCount` or `mutationCounts` functions.

loci A vector of loci indexes. Will be ignored only when single rate is specified. Default to all loci.

maxAllele Maximum allowable allele. Interpreted by each sub mutaor class. Default to `pop.maxAllele()`.

rate Can be a number (uniform rate) or an array of mutation rates (the same length as `loci`)

Member Functions

`x.apply(pop)` Apply a mutator

`x.clone()` Deep copy of a `mutator`

`x.maxAllele()` Return maximum allowable allele number

`x.mutate(allele)` Describe how to mutate a single allele

`x.mutationCount(locus)` Return mutation count at `locus`

`x.mutationCounts()` Return mutation counts

`x.rate()` Return the mutation rate

`x.setMaxAllele(maxAllele)` Set maximum allowable allele

`x.setRate(rate, loci=[])` Set an array of mutation rates

2.4.2 Class `kamMutator` (Function form: `KamMutate`)

K-Allele Model mutator.

Details

This mutator mutate an allele to another allelic state with equal probability. The specified mutation rate is actually the 'probability to mutate'. So the mutation rate to any other allelic state is actually $\frac{rate}{K-1}$, where K is specified by parameter `maxAllele`.

Initialization

Create a K-Allele Model mutator

```
kamMutator(rate=[], loci=[], maxAllele=0, output=">", outputExpr="",
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

loci A vector of loci indexes. Will be ignored only when single rate is specified. Default to all loci.

maxAllele Maximum allele that can be mutated to. For binary libraries allelic states will be `[0, maxAllele]`. Otherwise, they are `[1, maxAllele]`.

rate Mutation rate. It is the 'probability to mutate'. The actual mutation rate to any of the other `K-1` allelic states are `rates/(K-1)`.

Member Functions

`x.clone()` Deep copy of a `kamMutator`

`x.mutate(allele)` Mutate to a state other than current state with equal probability

Example

Example 2.6: Operator kamMutator

```
>>> simu = simulator(population(size=5, loci=[3,5]), noMating())
>>> simu.step([
...     kamMutator( rate=[.2,.6,.5], loci=[0,2,6], maxAllele=9),
...     dumper(alleleOnly=True)])
individual info:
sub population 0:
   0: MU   0  0  0   0  0  0  6  0 |   5  0  1   0  0  0  0  0
   1: MU   0  0  0   0  0  0  0  0 |   0  0  0   0  0  0  0  0
   2: MU   0  0  1   0  0  0  0  0 |   0  0  5   0  0  0  0  0
   3: MU   0  0  0   0  0  0  0  0 |   0  0  9   0  0  0  4  0
   4: MU   0  0  6   0  0  0  1  0 |   0  0  5   0  0  0  3  0
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.4.3 Class `smmMutator` (Function form: `SmmMutate`)

The stepwise mutation model.

Details

STEPWISE MUTATION MODEL (SMM) assumes that alleles are represented by integer values and that a mutation either increases or decreases the allele value by one. For variable number tandem repeats loci (VNTR), the allele value is generally taken as the number of tandem repeats in the DNA sequence.

Initialization

Create a SMM mutator

```
smmMutator(rate=[], loci=[], maxAllele=0, incProb=0.5, output=">",
outputExpr="", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

The stepwise mutation model (SMM) is developed for allozymes. It provides better description for these kinds of evolutionary processes. Please see `mutator` for the description of other parameters.

incProb Probability to increase allele state. Default to `0.5`.

Member Functions

`x.clone()` Deep copy of a `smmMutator`

Example

Example 2.7: Operator smmMutator

```
>>> simu = simulator(population(size=3, loci=[3,5]), noMating())
>>> simu.step([
...     initByFreq( [.2,.3,.5]),
...     smmMutator(rate=1,  incProb=.8),
...     dumper(alleleOnly=True, stage=PrePostMating)])
individual info:
sub population 0:
   0: FU   2  2  1   2  0  1  0  0 |   1  1  2   2  2  0  0  2
   1: FU   2  2  2   2  0  2  2  2 |   2  1  1   2  1  1  2  2
   2: MU   1  2  1   1  1  0  2  2 |   1  1  1   1  1  2  1  2
End of individual info.


No ancenstral population recorded.
individual info:
sub population 0:
   0: FU   1  3  2   3  0  0  1  1 |   2  2  3   3  3  1  1  3
   1: FU   1  3  3   1  1  1  3  3 |   3  2  2   3  0  0  3  1
   2: MU   2  1  2   2  2  1  3  1 |   2  2  2   0  2  3  0  3
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.4.4 Class `gsmMutator` (Function form: `GsmMutate`)

Generalized stepwise mutation model

Details

GENERALIZED STEPWISE MUTATION MODEL (GSM) is an extension to stepwise mutation model. This model assumes that alleles are represented by integer values and that a mutation either increases or decreases the allele value by a random value. In other words, in this model the change in the allelic state is drawn from a random distribution. A *geometric generalized stepwise model* uses a geometric distribution with parameter p, which has mean $\frac{p}{1-p}$ and variance $\frac{p}{(1-p)^2}$.

`gsmMutator` implements both models. If you specify a Python function without a parameter, this mutator will use its return value each time a mutation occur; otherwise, a parameter p should be provided and the mutator will act as a geometric generalized stepwise model.

Initialization

Create a `gsmMutator`

```
gsmMutator(rate=[], loci=[], maxAllele=0, incProb=0.5, p=0,
func=None, output=">", outputExpr="", stage=PostMating, begin=0,
end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

The generalized stepwise mutation model (GMM) is developed for allozymes. It provides better description for these kinds of evolutionary processes. Please see `mutator` for the description of other parameters.

func A function that returns the number of steps. This function does not accept any parameter.

incProb Probability to increase allele state. Default to `0.5`.

Member Functions

`x.clone()` Deep copy of a `gsmMutator`

`x.mutate(allele)` Mutate according to the GSM model

2.4.5 Class `pyMutator` (Function form: `PyMutate`)

A hybrid mutator.

Details

Parameters such as mutation rate of this operator are set just like others and you are supposed to provide a Python function to return a new allele state given an old state. `pyMutator` will choose an allele as usual and call your function to mutate it to another allele.

Initialization

Create a `pyMutator`

```
pyMutator(rate=[], loci=[], maxAllele=0, func=None, output=">",
outputExpr="", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

x.clone() Deep copy of a pyMutator

x.mutate(allele) Mutate according to the mixed model

Example

Example 2.8: Operator pyMutator

```
>>> def mut(x):
...   return 8
...
>>> simu = simulator(population(size=3, loci=[3,5]), noMating())
>>> simu.step([
...   pyMutator(rate=.5, loci=[3,4,5], func=mut),
...   dumper(alleleOnly=True)])
individual info:
sub population 0:
   0: MU   0  0  0   8  8  8  0  0 |   0  0  0   8  8  8  0  0
   1: MU   0  0  0   0  0  8  0  0 |   0  0  0   8  8  0  0  0
   2: MU   0  0  0   8  8  8  0  0 |   0  0  0   0  0  8  0  0
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.4.6 Class pointMutator (Function form: PointMutate)

Point mutator

Details

Mutate specified individuals at a specified loci to a spcified allele. I.e., this is a non-random mutator used to introduce diseases etc. pointMutator, as its name suggest, does point mutation. This mutator will turn alleles at loci on the first chromosome copy to toAllele for individualinds. You can specify atPloidy to mutate other, or all ploidy copies.

Initialization

Create a pointMutator

```
pointMutator(loci, toAllele, atPloidy=[], inds=[], output=">",
outputExpr="", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Please see mutator for the description of other parameters.

inds Individuals who will mutate

toAllele Allele that will be mutate to

Member Functions

x.apply(pop) Apply a pointMutator

x.clone() Deep copy of a `pointMutator`

x.mutationCount(locus) Return mutation count at `locus`

x.mutationCounts() Return mutation counts

2.5 Recombination

2.5.1 Class `recombinator`

Recombination

Details

In simuPOP, only one recombinator is provided. Recombination events between loci a/b and b/c are independent, otherwise there will be some linkage between loci, users need to specify physical recombination rate between adjacent loci. In addition, for the recombinator

- it only works for diploid (and for females in haplodiploid) populations.
- the recombination rate must be comprised between `0.0` and `0.5`. A recombination rate of `0.0` means that the loci are completely linked, and thus behave together as a single linked locus. A recombination rate of `0.5` is equivalent to free recombination. All other values between `0.0` and `0.5` will represent various linkage intensities between adjacent pairs of loci. The recombination rate is equivalent to `1-linkage` and represents the probability that the allele at the next locus is randomly drawn.

Initialization

Recombine chromosomes from parents

```
recombinator(intensity=-1, rate=[], afterLoci=[], maleIntensity=-1,
maleRate=[], maleAfterLoci=[], begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

afterLoci An array of locus indexes. Recombination will occur after these loci. If `rate` is also specified, they should have the same length. Default to all loci (but meaningless for those loci located at the end of a chromosome). If this parameter is given, it should be ordered, and can not include loci at the end of a chromosome.

intensity Intensity of recombination. The actually recombination rate between two loci is determined by `intensity*locus distance` between them.

maleAfterLoci If given, males will recombine at different locations.

maleIntensity Recombination intensity for male individuals. If given, parameter `intensity` will be considered as female intensity.

maleRate Recombination rate for male individuals. If given, parameter `rate` will be considered as female recombination rate.

rate Recombination rate regardless of locus distance after all `afterLoci`. It can also be an array of recombination rates. Should have the same length as `afterLoci` or `totNumOfLoci()`. The recombination rates are independent of locus distance.

Note

There is no recombination between sex chromosomes of male individuals if `sexChrom()=True`. This may change later if the exchanges of genes between pseudoautosomal regions of `XY` need to be modeled.

Member Functions

`x.clone()` Deep copy of a recombinator

`x.recCount(locus)` Return recombination count

`x.recCounts()` Return recombination counts

Example

Example 2.9: Operator recombinator

```
>>> simu = simulator(
...     population(4, loci=[4,5,6], maxAllele=9,
...     infoFields=['father_idx', 'mother_idx']),
...     randomMating())
>>> simu.step([
...   parentsTagger(),
...   ],
...   preOps = [initByFreq([.2,.2,.4,.2]), dumper(alleleOnly=True) ],
...   postOps = [ dumper(alleleOnly=True)]
... )
individual info:
sub population 0:
   0: FU 0022 02222 302110 | 0330 02323 221121
   1: MU 0211 03122 221322 | 1133 03222 303311
   2: MU 2323 22022 210223 | 2222 12330 302220
   3: MU 3020 13223 012002 | 3022 13020 133202
End of individual info.


No ancenstral population recorded.
individual info:
sub population 0:
   0: MU 3022 13223 012002 | 0022 02222 221121
   1: FU 3020 13020 133202 | 0330 02222 221121
   2: MU 2222 12330 210223 | 0022 02323 302110
   3: FU 1133 03122 221322 | 0330 02323 221121
End of individual info.


No ancenstral population recorded.
True
>>> simu.step([
...   parentsTagger(),
...   recombinator(rate=[1,1,1], afterLoci=[2,6,10])
...   ],
...   postOps = [ dumper(alleleOnly=True)]
... )
individual info:
sub population 0:
```

```
   0: MU 3020 13022 131121 | 0022 02223 011121
   1: FU 0333 03123 221322 | 3022 02223 011121
   2: FU 0333 03123 221121 | 0022 12323 300223
   3: FU 0330 13022 131121 | 0022 13222 011121
End of individual info.


No ancenstral population recorded.
True
>>>
```

2.6 Selection

2.6.1 Mechanism

It is not very clear that our method agrees with the traditional 'average number of offspring' definition of fitness. (Note that this concept is very difficult to simulate because we do not know who will determine the number of offspring if two parents are involved.) We can, instead, look at the consequence of selection in a simple case (as derived in any population genetics textbook):

> At generation t, genotype P_{11}, P_{12}, P_{22} has fitness values $w_{11}, w_{12,}w_{22}$ respectively. In the next generation the proportion of genotype P_{11} etc., should be

$$\frac{P_{11}w_{11}}{P_{11}w_{11} + P_{12}w_{12} + P_{22}w_{22}}$$

Now, using the 'ability-to-mate' approach, for the sexless case, the proportion of genotype 11 will be the number of 11 individuals times its probability to be chosen:

$$n_{11}\frac{w_{11}}{\sum_{n=1}^{N} w_n}$$

This is, however, exactly

$$n_{11}\frac{w_{11}}{\sum_{n=1}^{N} w_n} = n_{11}\frac{w_{11}}{n_{11}w_{11} + n_{12}w_{12} + n_{22}w_{22}} = \frac{P_{11}w_{11}}{P_{11}w_{11} + P_{12}w_{12} + P_{22}w_{22}}$$

The same argument applies to the case of arbitrary number of genotypes and random mating.

The following operators, when applied, will set a variable `fitness` and an indicator so that selector-aware mating scheme can select individuals according to these values. This has two consequences:

- selector alone can not do selection! Only mating schemes can actually select on individuals.
- selector has to be `PreMating` operator. This is not a problem when you use the operator form of the selectors since their default stage is `PreMating`. However, if you use the function form of these selectors in a `pyOperator`, make sure to set the stage of `pyOperator` to `PreMating`.

2.6.2 Class `selector`

A base selection operator for all selectors.

Details

Genetic selection is tricky to simulate since there are many different *fitness* values and many different ways to apply selection. `simuPOP` employs an *'ability-to-mate'* approach. Namely, the probability that an individual will be chosen for mating is proportional to its fitness value. More specifically,

- `PreMating` selectors assign fitness values to each individual, and mark part or all subpopulations as under selection.
- During sexless mating (e.g. `binomialSelection` mating scheme), individuals are chosen at probabilities that are proportional to their fitness values. If there are N individuals with fitness values $f_i, i = 1, ..., N$, individual i will have probability $\frac{f_i}{\sum_j f_j}$ to be chosen and passed to the next generation.
- During `randomMating`, males and females are separated. They are chosen from their respective groups in the same manner as `binomialSelection` and mate.

All of the selection operators, when applied, will set an information field `fitness` (configurable) and then mark part or all subpopulations as under selection. (You can use different selectors to simulate varying selection intensity for different subpopulations). Then, a 'selector-aware' mating scheme can select individuals according to this `fitness` information field. This implies that

- Only mating schemes can actually select individuals.
- Selector has to be `PreMating` operator. This is not a problem when you use the operator form of the selectors since their default stage is `PreMating`. However, if you use the function form of these selectors in a `pyOperator`, make sure to set the stage of `pyOperator` to `PreMating`.

Note

You can not apply two selectors to the same subpopulation, because only one fitness value is allowed for each individual.

Initialization

Create a selector

```
selector(subPops=[], stage=PreMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=["fitness"])
```

subPop A shortcut to `subPops=[subPop]`

subPops Subpopulations that the selector will apply to. Default to all.

Member Functions

`x.apply(pop)` Set fitness to all individuals. No selection will happen!

`x.clone()` Deep copy of a selector

2.6.3 Class `mapSelector` (Function form: `MapSelector`)

Selection according to the genotype at one locus

Details

This map selector implements selection at one locus. A user provided dictionary (map) of genotypes will be used in this selector to set each individual's fitness value.

Initialization

Create a map selector

```
mapSelector(loci, fitness, phase=False, subPops=[], stage=PreMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=["fitness"])
```

fitness A dictionary of fitness values. The genotype must be in the form of `'a-b'` for a single locus, and `'a-b|c-d|e-f'` for multi-loci.

loci The locus indexes. The genotypes at these loci will be used to determine fitness value.

locus The locus index. A shortcut to `loci=[locus]`

output And other parameters please refer to help(baseOperator.__init__)

phase If `True`, genotypes `a-b` and `b-a` will have different fitness values. Default to false.

Member Functions

`x.clone()` Deep copy of a map selector

Example

Example 2.10: Use of mapSelector

```
>>> simu = simulator(
...     population(size=1000, ploidy=2, loci=[1], infoFields=['fitness']),
...     randomMating())
>>> s1 = .1
>>> s2 = .2
>>> simu.evolve([
...     stat( alleleFreq=[0], genoFreq=[0]),
...     mapSelector(locus=0, fitness={'0-0':(1-s1), '0-1':1, '1-1':(1-s2)}),
...     pyEval(r"'%.4f\n' % alleleFreq[0][1]", step=100)
...     ],
...     preOps=[  initByFreq(alleleFreq=[.2,.8])],
...     end=300)
0.7740
0.3310
0.3635
0.3335
True
>>>
```

The example for `class mapSelector` is a typical example of heterozygote superiority. When $w_{11} < w_{12} > w_{22}$, the genotype frequencies will go to an equilibrium state. Theoretically, if

$$\begin{aligned} s_1 &= w_{12} - w_{11} \\ s_2 &= w_{12} - w_{22} \end{aligned}$$

the stable allele frequency of allele 1 is

$$p = \frac{s_2}{s_1 + s_2}$$

Which is .677 in the example ($s_1 = .1$, $s_2 = .2$).

2.6.4 Class `maSelector` (Function form: `MaSelect`)

Multiple allele selector (selection according to wildtype or diseased alleles)

Details

This is called 'multiple-allele' selector. It separate alleles into two groups: wildtype and disease alleles. Wildtype alleles are specified by parameter `wildtype` and any other alleles are considered as diseased alleles.This selector accepts an array of fitness values:

- For single-locus, `fitness` is the fitness for genotype AA, Aa, aa, while A stands for wildtype alleles.
- For a two-locus model, `fitness` is the fitness for genotype AABB, AABb, AAbb, AaBB, AbBb, Aabb, aaBB, aaBb and aaBb.
- For a model with more than two loci, use a table of length $3\hat{}\{n\}$ in a order similar to the two-locus model.

Initialization

Create a multiple allele selector

```
maSelector(loci, fitness, wildtype, subPops=[], stage=PreMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=["fitness"])
```

Please refer to `basicSelector` for other parameter descriptions.

fitness For the single locus case, fitness is an array of fitness of AA, Aa, aa. A is the wildtype group. In the case of multiple loci, fitness should be in the order of AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb, aabb.

output And other parameters please refer to help(baseOperator.__init__)

wildtype An array of alleles in the wildtype group. Any other alleles are considered to be diseased alleles. Default to `[0]`.

Note

- `maSelector` only works for diploid populations now.
- `wildtype` at all loci are the same.

Member Functions

`x.clone()` Deep copy of a `maSelector`

Example

Example 2.11: Use of maSelector

```
>>> simu = simulator(
...     population(size=1000, ploidy=2, loci=[1], infoFields=['fitness']),
...     randomMating())
>>> s1 = .1
>>> s2 = .2
>>> simu.evolve(
```

```
...     preOps=[initByFreq(alleleFreq=[.2,.8])],
...     ops = [
...         stat( alleleFreq=[0], genoFreq=[0]),
...         maSelector(locus=0, fitness=[1-s1, 1, 1-s2]),
...         pyEval(r"'%.4f\n' % alleleFreq[0][1]", step=100)
...     ],
...     end=300)
0.7915
0.3210
0.3645
0.2865
True
>>>
```

2.6.5 Class `mlSelector` (Function form: `MlSelect`)

Selection according to genotypes at multiple loci in a multiplicative model

Details

This selector is a 'multiple-loci model' selector. The selector takes a vector of selectors (can not be another `mlSelector`) and evaluate the fitness of an individual as the the product or sum of individual fitness values. The mode is determined by parameter `mode`, which takes the value

- `SEL_Multiplicative`: the fitness is calculated as $f = \prod_i f_i$.
- `SEL_Additive`: the fitness is calculated as $f = \max(0, 1 - \sum_i (1 - f_i)) = \max(0, 1 - \sum_i s_i)$. f will be set to `0` when $f < 0$. In this case, s_i are added, not f_i directly.

Initialization

Create a multi-loci selector

```
mlSelector(selectors, mode=SEL_Multiplicative, subPops=[],
stage=PreMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=["fitness"])
```

Please refer to `mapSelector` for other parameter descriptions.

selectors A list of selectors

Member Functions

`x.clone()` Deep copy of a `mlSelector`

2.6.6 Class `pySelector` (Function form: `PySelect`)

Selection using user provided function

Details

`PYSELECTOR` assigns fitness values by calling a user provided function. It accepts a list of susceptibility loci and a Python function. For each individual, this operator will pass the genotypes at these loci and the generation number

and use the returned value as the fitness value. The genotypes are arranged in the order of `0-0,0-1,1-0,1-1` etc. where X-Y represents locus X - ploidy Y.

Initialization

Create a Python hybrid selector

```
pySelector(loci, func, subPops=[], stage=PreMating, begin=0, end=-1,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=["fitness"])
```

func A Python function that accepts genotypes at susceptibility loci generation number, and return fitness value.

loci Susceptibility loci. The genotype at these loci will be passed to `func`.

output And other parameters please refer to help(baseOperator.__init__)

Member Functions

`x.clone()` Deep copy of a `pySelector`

2.7 Penetrance

2.7.1 Class `penetrance`

Base class of all penetrance operators.

Details

Penetrance is the probability that one will have the disease when he has certain genotype(s). Calculation and the parameter set of penetrance are similar to those of fitness. An individual will be randomly marked as affected/unaffected according to his/her penetrance value. For example, an individual will have probability 0.8 to be affected if the penetrance is 0.8.

Penetrance can be applied at any stage (default to `DuringMating`). When a penetrance operator is applied, it calculate the penetrance value of each offspring and assign affected status accordingly. Penetrance can also be used `PreMating` or `PostMating`. In these cases, the affected status will be set to all individuals according to their penetrance values.Pentrance values are used to set the affectedness of individuals, and are usually not saved. If you would like to know the penetrance value, you need to

- use `addInfoField('penetrance')` to the population to analyze. (Or use `infoFields` parameter of the population constructor), and
- use e.g., `mlPenetrance(...., infoFields=['penetrance'])` to add the penetrance field to the penetrance operator you use. You may choose a name other than `'penetrance'` as long as the field names for the operator and population match.

Penetrance functions can be applied to the current, all, or certain number of ancestral generations. This is controlled by the `ancestralGen` parameter, which is default to `-1` (all available ancestral generations). You can set it to `0` if you only need affection status for the current generation, or specify a number `n` for the number of ancestral generations (n + 1 total generations) to process. Note that `ancestralGen` parameter is ignored if the penetrance operator is used as a during mating operator.

Initialization

Create a penetrance operator

```
penetrance(ancestralGen=-1, stage=DuringMating, begin=0, end=-1,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

default to be always active.

ancestralGen If this parameter is set to be `0`, then apply penetrance to the current generation; if `-1`, apply to all generations; otherwise, apply to the specified number of ancestral generations

infoFields If one field is specified, it will be used to store penetrance values.

stage Specify the stage this operator will be applied, default to `DuringMating`.

Member Functions

`x.apply(pop)` Set penetrance to all individuals and record penetrance if requested

`x.clone()` Deep copy of a penetrance operator

2.7.2 Class `mapPenetrance` (Function form: `MapPenetrance`)

Penetrance according to the genotype at one locus

Details

Assign penetrance using a table with keys `'X-Y'` where `X` and `Y` are allele numbers.

Initialization

Create a map penetrance operator

```
mapPenetrance(loci, penet, phase=False, ancestralGen=-1,
stage=DuringMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

loci The loci indexes. The genotypes of these loci will be used to determine penetrance.

locus The locus index. Shortcut to `loci=[locus]`

output And other parameters please refer to help(baseOperator.__init__)

penetrance A dictionary of penetrance. The genotype must be in the form of 'a-b' for a single locus.

phase If True, `a/b` and `b/a` will have different penetrance values. Default to `False`.

Member Functions

`x.clone()` Deep copy of a map penetrance operator

2.7.3 Class `maPenetrance` (Function form: `MaPenetrance`)

Multiple allele penetrance operator

Details

This is called 'multiple-allele' penetrance. It separates alleles into two groups: wildtype and disease alleles. Wildtype alleles are specified by parameter `wildtype` and any other alleles are considered as diseased alleles.

`maPenetrance` accepts an array of fitness for AA, Aa, aa in the single-locus case, and a longer table for multi-locus case. Penetrance is then set for any given genotype.

Initialization

Create a multiple allele penetrance operator (penetrance according to diseased or wildtype alleles)

```
maPenetrance(loci, penet, wildtype, ancestralGen=-1,
stage=DuringMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

loci The loci indexes. The genotypes of these loci will be examed.

locus The locus index. The genotype of this locus will be used to determine penetrance.

output And other parameters please refer to help(baseOperator.__init__)

penetrance An array of penetrance values of AA, Aa, aa. A is the wild type group. In the case of multiple loci, fitness should be in the order of AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb, aabb.

wildtype An array of alleles in the wildtype group. Any other alleles will be considered as in the disease allele group.

Member Functions

`x.clone()` Deep copy of a multi-allele penetrance operator

2.7.4 Class `mlPenetrance` (Function form: `MlPenetrance`)

Penetrance according to the genotype according to a multiple loci multiplicative model

Details

`MLPENTRANCE` is the 'multiple-locus' penetrnace calculator. It accepts a list of penetrances and combine them according to the `mode` parameter, which takes one of the following values:

- `PEN_Multiplicative:` the penetrance is calculated as $f = \prod f_i$.
- `PEN_Additive:` the penetrance is calculated as $f = \min(1, \sum f_i)$. f will be set to `1` when $f < 0$. In this case, s_i are added, not f_i directly.
- `PEN_Heterogeneity:` the penetrance is calculated as $f = 1 - \prod (1 - f_i)$.

Please refer to Neil Risch (1990) for detailed information about these models.

Initialization

Create a multiple loci penetrance operator using a multiplicative model

```
mlPenetrance(peneOps, mode=PEN_Multiplicative, ancestralGen=-1,
stage=DuringMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

mode Can be one of `PEN_Multiplicative`, `PEN_Additive`, and `PEN_Heterogeneity`

peneOps A list of penetrance operators

Member Functions

`x.clone()` Deep copy of a multi-loci penetrance operator

Example

Example 2.12: Use of multi-locus penetrance operator

```
>>> pop = population(1000, loci=[3])
>>> InitByFreq(pop, [0.3, 0.7])
>>> pen = []
>>> for loc in (0, 1, 2):
...     pen.append( maPenetrance(locus=loc, wildtype=[1],
...         penetrance=[0, 0.3, 0.6] ) )
...
>>> # the multi-loci penetrance
>>> MlPenetrance(pop, mode=PEN_Multiplicative, peneOps=pen)
>>> Stat(pop, numOfAffected=True)
>>> print pop.dvars().numOfAffected
8
>>>
```

2.7.5 Class `pyPenetrance` (Function form: `PyPenetrance`)

Assign penetrance values by calling a user provided function

Details

For each individual, users provide a function to calculate penetrance. This method is very flexible but will be slower than previous operators since a function will be called for each individual.

Initialization

Provide locus and penetrance for 11, 12, 13 (in the form of dictionary)

```
pyPenetrance(loci, func, ancestralGen=-1, stage=DuringMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

func A user-defined Python function that accepts an array of genotypes at susceptibility loci and return a penetrance value. The returned value should be between `0` and `1`.

loci Disease susceptibility loci. The genotypes at these loci will be passed to the provided Python function in the form of `loc1_1, loc1_2, loc2_1, loc2_2, ...` if the individuals are diploid.

output And other parameters please refer to help(baseOperator.__init__)

Member Functions

`x.clone()` Deep copy of a Python penetrance operator

Example

Example 2.13: Use of python penetrance operator

```
>>> pop = population(1000, loci=[3])
>>> InitByFreq(pop, [0.3, 0.7])
>>> def peneFunc(geno):
...     p = 1
...     for l in range(len(geno)/2):
...         p *= (geno[l*2]+geno[l*2+1])*0.3
...     return p
...
>>> PyPenetrance(pop, func=peneFunc, loci=(0, 1, 2))
>>> Stat(pop, numOfAffected=True)
>>> print pop.dvars().numOfAffected
75
>>> #
>>> # You can also define a function, that returns a penetrance
>>> # function using given parameters
>>> def peneFunc(table):
...     def func(geno):
...       return table[geno[0]][geno[1]]
...     return func
...
>>> # then, given a table, you can do
>>> PyPenetrance(pop, loci=(0, 1, 2),
...     func=peneFunc( ((0,0.5),(0.3,0.8)) ) )
>>>
```

2.8 Quantitative Trait

2.8.1 Class `quanTrait`

Base class of quantitative trait

Details

Quantitative trait is the measure of certain phenotype for given genotype. Quantitative trait is similar to penetrance in that the consequence of penetrance is binary: affected or unaffected; while it is continuous for quantitative trait.

In simuPOP, different operators/functions were implemented to calculate quantitative traits for each individual and store the values in the information fields specified by user (default to `qtrait`). The quantitative trait operators also accept the `ancestralGen` parameter to control the number of generations for which the `qtrait` information field will be set.

Initialization

Create a quantitative trait operator, default to be always active

```
quanTrait(ancestralGen=-1, stage=PostMating, begin=0, end=-1, step=1,
at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=["qtrait"])
```

Member Functions

`x.apply(pop)` Set `qtrait` to all individual

`x.clone()` Deep copy of a quantitative trait operator

2.8.2 Class `mapQuanTrait` (Function form: `MapQuanTrait`)

Quantitative trait according to genotype at one locus

Details

Assign quantitative trait using a table with keys 'X-Y' where X and Y are allele numbers. If parameter `sigma` is not zero, the returned value is the sum of the trait plus $N\left(0,\sigma^2\right)$. This random part is usually considered as the environmental factor of the trait.

Initialization

Create a map quantitative trait operator

```
mapQuanTrait(loci, qtrait, sigma=0, phase=False, ancestralGen=-1,
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=["qtrait"])
```

loci An array of locus indexes. The quantitative trait is determined by genotype at these loci.

locus The locus index. The quantitative trait is determined by genotype at this locus.

output And other parameters please refer to help(baseOperator.__init__)

phase If `True`, a/b and b/a will have different quantitative trait values. Default to False.

qtrait A dictionary of quantitative traits. The genotype must be in the form of 'a-b'. This is the mean of the quantitative trait. The actual trait value will be `N(mean, sigma^2)`. For multiple loci, the form is 'a-b|c-d|e-f' etc.

sigma Standard deviation of the environmental factor `N(0, sigma^2)`.

Member Functions

`x.clone()` Deep copy of a map quantitative trait operator

2.8.3 Class `maQuanTrait` (Function form: `MaQuanTrait`)

Multiple allele quantitative trait (quantitative trait according to disease or wildtype alleles)

Details

This is called 'multiple-allele' quantitative trait. It separates alleles into two groups: wildtype and disease susceptibility alleles. Wildtype alleles are specified by parameter `wildtype` and any other alleles are considered as disease alleles. `maQuanTrait` accepts an array of fitness. Quantitative trait is then set for any given genotype. A standard normal distribution $N\left(0,\sigma^2\right)$ will be added to the returned trait value.

Initialization

Create a multiple allele quantitative trait operator

```
maQuanTrait(loci, qtrait, wildtype, sigma=[], ancestralGen=-1,
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=["qtrait"])
```

Please refer to `quanTrait` for other parameter descriptions.

output And other parameters please refer to help(baseOperator.__init__)

qtrait An array of quantitative traits of AA, Aa, aa. A is the wild type group

sigma An array of standard deviations for each of the trait genotype (AA, Aa, aa)

wildtype An array of alleles in the wildtype group. Any other alleles will be considered as disease alleles. Default to `[0]`.

Member Functions

`x.clone()` Deep copy of a multiple allele quantitative trait

2.8.4 Class `mlQuanTrait` (Function form: `MlQuanTrait`)

Quantitative trait according to genotypes from a multiple loci multiplicative model

Details

`MLQUANTRAIT` is a 'multiple-loci' quantitative trait calculator. It accepts a list of quantitative traits and combine them according to the `mode` parameter, which takes one of the following values

- `QT_Multiplicative:` the mean of the quantitative trait is calculated as $f = \prod f_i$.
- `QT_Additive:` the mean of the quantitative trait is calculated as $f = \sum f_i$.

Note that all σ_i (for f_i) and σ (for f) will all be considered. I.e, the trait value should be

$$f = \sum_i \left(f_i + N\left(0, \sigma_i^2\right)\right) + \sigma^2$$

for `QT_Additive` case. If this is not desired, you can set some of the σ to zero.

Initialization

Multiple loci quantitative trait using a multiplicative model

```
mlQuanTrait(qtraits, mode=QT_Multiplicative, sigma=0,
ancestralGen=-1, stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=["qtrait"])
```

Please refer to `quanTrait` for other parameter descriptions.

mode Can be one of `QT_Multiplicative` and `QT_Additive`

qtraits A list of quantitative traits

Member Functions

`x.clone()` Deep copy of a multiple loci quantitative trait operator

2.8.5 Class `pyQuanTrait` (Function form: `PyQuanTrait`)

Quantitative trait using a user provided function

Details

For each individual, a user provided function is used to calculate quantitative trait.

Initialization

Create a Python quantitative trait operator

```
pyQuanTrait(loci, func, ancestralGen=-1, stage=PostMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=["qtrait"])
```

Please refer to `quanTrait` for other parameter descriptions.

func A Python function that accepts genotypes at susceptibility loci and returns the quantitative trait value.

loci Susceptibility loci. The genotypes at these loci will be passed to `func`.

output And other parameters please refer to help(baseOperator.__init__)

Member Functions

`x.clone()` Deep copy of a Python quantitative trait operator

2.9 Ascertainment

2.9.1 Class `sample`

Base class of other sample operator

Details

Ascertainment/sampling refers to ways to select individuals from a population. In simuPOP, ascerntainment operators form separate populations in a population's namespace. All the ascertainment operators work like this except for `pySubset` which shrink the population itself.

Individuals in sampled populations may or may not keep their original order but their indexes in the whole population are stored in a information field `oldindex`. That is to say, you can use `ind.info('oldindex')` to check the original position of an individual.

Two forms of sample size specification are supported: with or without subpopulation structure. For example, the `size` parameter of `randomSample` can be a number or an array (which has the length of the number of subpopulations). If a number is given, a sample will be drawn from the whole population, regardless of the population structure. If an array is given, individuals will be drawn from each subpopulation `sp` according to `size[sp]`.

An important special case of sample size specification occurs when `size=[]` (default). In this case, usually all qualified individuals will be returned.

The function forms of these operators are a little different from others. They do return a value: an array of samples.

Initialization

Draw a sample

```
sample(name="sample", nameExpr="", times=1, saveAs="", saveAsExpr="",
format="auto", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Please refer to `baseOperator::__init__` for other parameters.

format Format to save the samples

name Name of the sample in local namespace. This variable is an array of populations of size `times`. Default to `sample`. If `name=''` is set, samples will not be saved in local namespace.

nameExpr Expression version of parameter `name`. If both `name` and `nameExpr` are empty, do not store `pop`. This expression will be evaluated dynamically in population's local namespace.

saveAs Filename to save the samples

saveAsExpr Expression version of parameter `saveAs`. It will be evaluated dynamically in population's local namespace.

times How many times to sample from the population. This is usually `1`, but we may want to take several random samples.

Member Functions

`x.apply(pop)` Apply the `sample` operator

`x.clone()` Deep copy of a `sample` operator

`x.samples(pop)` Return the samples

2.9.2 Class `pySubset` (Function form: `PySubset`)

Shrink population

Details

This operator shrinks a population according to a given array or the `subPopID()` value of each indvidual. individuals with negative subPop ID are removed.

Initialization

Create a `pySubset` operator

```
pySubset(keep=[], stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

keep An array of subpopulation IDs for each individual.

Member Functions

`x.apply(pop)` Apply the `pySubset` operator

`x.clone()` Deep copy of a `pySubset` operator

2.9.3 Class `pySample` (Function form: `PySample`)

Python sampler.

Details

A Python sampler that generate a sample with given individuals.

Initialization

Create a Python sampler

```
pySample(keep, keepAncestralPops=-1, name="sample", nameExpr="",
times=1, saveAs="", saveAsExpr="", format="auto", stage=PostMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

This sampler accepts a Python array which will be assigned to each individual as subPOP ID. Individuals with positive subPOPID will then be picked out and form a sample. Please refer to class `sample` for other parameter descriptions.

keep Subpopulation IDs of all individuals

keepAncestralPop The number of ancestral populations that will be kept. If `-1`, keep all ancestral populations (default). If `0`, no ancestral population will be kept.

Member Functions

`x.clone()` Deep copy of a Python sampler

`x.drawsample(pop)` Draw a Python sample

2.9.4 Class `randomSample` (Function form: `RandomSample`)

Randomly draw a sample from a population

Details

This operator will randomly choose `size` individuals (or `size[i]` individuals from subpopulation `i`) and return a new population. The function form of this operator returns the samples directly. The operator keeps samples in an array `name` in the local namespace. You may access them through `dvars()` or `vars()` functions.

The original subpopulation structure/boundary is kept in the samples.

Initialization

Draw a random sample, regardless of the affected status

```
randomSample(size=[], name="sample", nameExpr="", times=1, saveAs="",
saveAsExpr="", format="auto", stage=PostMating, begin=0, end=-1,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Please refer to class `sample` for other parameter descriptions.

size Size of the sample. It can be either a number which represents the overall sample size, regardless of the population structure; or an array which represents the number of samples drawn from each subpopulation.

Note

Ancestral populations will not be copied to the samples.

Member Functions

`x.clone()` Deep copy of a `randomSample` operator

2.9.5 Class `caseControlSample` (Function form: `CaseControlSample`)

Draw a case-control sample from a population

Details

This operator will randomly choose `cases` affected individuals and `controls` unaffected individuals as a sample. The affected status is usually set by penetrance functions/operators. The sample populations will have two subpopulations: cases and controls.

You may specify the number of cases and the number of controls from each subpopulation using the array form of the parameters. The sample population will still have only two subpoulations (cases/controls) though.

A special case of this sampling scheme occurs when one of or both `cases` and `controls` are omitted (zeros). In this case, all cases and/or controls are chosen. If both parameters are omitted, the sample is effectively the same population with affected and unaffected individuals separated into two subpopulations.

Initialization

Draw cases and controls as a sample

```
caseControlSample(cases=[], controls=[], spSample=False,
name="sample", nameExpr="", times=1, saveAs="", saveAsExpr="",
format="auto", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Please refer to class `sample` for other parameter descriptions.

cases The number of cases, or an array of the numbers of cases from each subpopulation

controls The number of controls, or an array of the numbers of controls from each subpopulation

Member Functions

`x.clone()` Deep copy of a `caseControlSample` operator

2.9.6 Class `affectedSibpairSample` (Function form: `AffectedSibpairSample`)

Draw an affected sibling pair sample

Details

Special preparation for the population is needed in order to use this operator. Obviously, to obtain affected sibling pairs, we need to know the parents and the affectedness status of each individual. Furthermore, to get parental genotype, the population should have `ancestralDepth` at least `1`. The most important problem, however, comes from the mating scheme we are using.

`randomMating()` is usually used for diploid populations. The *realrandom* mating requires that a mating will generate only one offspring. Since parents are chosen with replacement, a parent can have multiple offspring with different parents. On the other hand, it is very unlikely that two offspring will have the same parents. The probability of having a sibling for an offspring is $\frac{1}{N^2}$ (if do not consider selection). Therefore, we will have to allow multiple offspring per mating at the cost of small effective population size.

All these requirements come at a cost: multiple ancestral populations, determining affectedness status and tagging will slow down evolution; multiple offspring will reduce effective population size. Fortunately, simuPOP is flexible enough to let all these happen only at the last several generations.

Initialization

Draw an affected sibling pair sample

```
affectedSibpairSample(size=[], chooseUnaffected=False,
countOnly=False, name="sample", nameExpr="", times=1, saveAs="",
saveAsExpr="", format="auto", stage=PostMating, begin=0, end=-1,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=["father_idx",
"mother_idx"])
```

Please refer to class `sample` for other parameter descriptions.

chooseUnaffected Instead of affected sibpairs, choose unaffected families.

countOnly Set variables about number of affected sibpairs, do not actually draw the sample

size The number of affected sibling pairs to be sampled. Can be a number or an array. If a number is given, it is the total number of sibpairs, ignoring population structure. Otherwise, given number of sibpairs are sampled from subpopulations. If size is unspecified, this operator will return all affected sibpairs.

Member Functions

`x.clone()` Deep copy of a `affectedSibpairSample` operator

`x.drawsample(pop)` Draw a sample

`x.prepareSample(pop)` Preparation before drawing a sample

2.9.7 Class `largePedigreeSample`

Draw a large pedigree sample

Initialization

Draw a large pedigree sample

```
largePedigreeSample(size=[], minTotalSize=0, maxOffspring=5,
minPedSize=5, minAffected=0, countOnly=False, name="sample",
nameExpr="", times=1, saveAs="", saveAsExpr="", format="auto",
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=["father_idx", "mother_idx"])
```

Please refer to class `sample` for other parameter descriptions.

countOnly Set variables about number of affected sibpairs, do not actually draw the sample.

maxOffspring The maximum number of offspring a parent may have

minAffected Minimal number of affected individuals in each pedigree, default to `0`

minPedSize Minimal pedigree size, default to `5`

minTotalSize The minimum number of individuals in the sample

Member Functions

`x.clone()` Deep copy of a `largePedigreeSample` operator

`x.drawsample(pop)` Draw a a large pedigree sample

`x.prepareSample(pop)` Preparation before drawing a sample

2.9.8 Class `nuclearFamilySample`

Draw a nuclear family sample

Initialization

Draw a nuclear family sample

```
nuclearFamilySample(size=[], minTotalSize=0, maxOffspring=5,
minPedSize=5, minAffected=0, countOnly=False, name="sample",
nameExpr="", times=1, saveAs="", saveAsExpr="", format="auto",
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=["father_idx", "mother_idx"])
```

Please refer to class `sample` for parameter descriptions.

Member Functions

`x.clone()` Deep copy of a `nuclearFamilySample` operator

`x.drawsample(pop)` Draw a nuclear family sample

`x.prepareSample(pop)` Preparation before drawing a sample

2.10 Statistics Calculation

2.10.1 Class `stator`

Base class of all the statistics calculator

Details

Operator `stator` calculate various basic statistics for the population and set variables in the local namespace. Other operators/functions can refer to the results from the namespace after `stat` is applied.

Initialization

Create a stator

```
stator(output="", outputExpr="", stage=PostMating, begin=0, end=-1,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.clone()` Deep copy of a stator

2.10.2 Class `stat` (Function form: `Stat`)

Calculate statistics

Details

Operator `stat` calculatse various basic statistics for the population and sets variables in the local namespace. Other operators/functions can refer to the results from the namespace after `stat` is applied. `Stat` is the function form of the operator.

Note that these statistics are dependent to each other. For example, heterotype and allele frequencies of related loci will be automatically calculated if linkage diseqilibrium is requested.

Initialization

Create an `stat` operator

```
stat(popSize=False, numOfMale=False, numOfMale_param={},
numOfAffected=False, numOfAffected_param={}, numOfAlleles=[],
numOfAlleles_param={}, alleleFreq=[], alleleFreq_param={},
heteroFreq=[], expHetero=[], expHetero_param={}, homoFreq=[],
genoFreq=[], haploFreq=[], LD=[], LD_param={}, association=[],
association_param={}, Fst=[], Fst_param={}, relGroups=[], relLoci=[],
rel_param={}, relBySubPop=False, relMethod=[], relMinScored=10,
hasPhase=False, midValues=False, output="", outputExpr="",
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

If only one item is specified, the outer `[]` can be ignored. I.e., `LD=[loc1, loc2]` is acceptable. This parameter will set the following variables. Please note that the difference between the data structures used for `ld` and `LD`. The names are potentially very confusing but I have no better idea.

- `ld['loc1-loc2']['allele1-allele2']`, `subPop[sp]['ld']['loc1-loc2']['allele1-allele2']`
- `ld_prime['loc1-loc2']['allele1-allele2']`, `subPop[sp]['ld_prime']['loc1-loc2']['allele1-a`
- `r2['loc1-loc2']['allele1-allele2']`, `subPop[sp]['r2']['loc1-loc2']['allele1-allele2']`
- `LD[loc1][loc2]`, `subPop[sp]['LD'][loc1][loc2]`
- `LD_prime[loc1][loc2]`, `subPop[sp]['LD_prime'][loc1][loc2]`
- `R2[loc1][loc2]`, `subPop[sp]['R2'][loc1][loc2]`

Fst Calculate F_{st}, F_{is}, F_{it}. For example, `Fst = [0,1,2]` will calculate F_{st}, F_{is}, F_{it} based on alleles at loci `0`, `1`, `2`. The locus-specific values will be used to calculate `AvgFst`, which is an average value over all alleles (Weir & Cockerham, 1984). Terms and values that match Weir & Cockerham:

- F (F_{IT}) the correlation of genes within individuals (inbreeding);
- θ (F_{ST}) the correlation of genes of difference individuals in the same population (will evaluate for each subpopulation and the whole population)
- f (F_{IS}) the correlation of genes within individuals within populations.

This parameter will set the following variables:

- `Fst[loc]`, `Fis[loc]`, `Fit[loc]`
- `AvgFst`, `AvgFis`, `AvgFit`.

Fst_param A dictionary of parameters of `Fst` statistics. Can be one or more items choosen from the following options: `Fst`, `Fis`, `Fit`, `AvgFst`, `AvgFis`, and `AvgFit`.

LD Calculate linkage disequilibria LD, LD' and r^2, given `LD=[ [loc1, loc2], [ loc1, loc2, allele1, allele2], ... ]` For each item `[loc1, loc2, allele1, allele2]`, D, D' and r^2 will be calculated based on `allele1` at `loc1` and `allele2` at `loc2`. If only two loci are given, the LD values are averaged over all allele pairs. For example, for allele A at locus `1` and allele B at locus `2`,

$$D = P_{AB} - P_A P_B$$

$$D' = D/D_{max}$$

$$D_{max} = \min\left(P_A\left(1-P_B\right), \left(1-P_A\right)P_B\right) \text{ if } D > 0 \min\left(P_A P_B, \left(1-P_A\right)\left(1-P_B\right)\right) \text{ if } D < 0$$

$$r^2 = \frac{D^2}{P_A\left(1-P_A\right)P_B\left(1-P_B\right)}$$

LD_param A dictionary of parameters of `LD` statistics. Can have key `stat` which is a list of statistics to calculate. Default to all. If any statistics is specified, only those specified will be calculated. For example, you may use `LD_param={LD_prime}` to calculate D' only, where `LD_prime` is a shortcut for `'stat':['LD_prime']`. Other parameters that you may use are:

- `subPop`, whether or not calculate statistics for subpopulations
- `midValues`, whether or not keep intermediate results.

alleleFreq An array of loci at which all allele frequencies will be calculated (`alleleFreq=[loc1, loc2, ...]` where `loc1` etc. are loci where allele frequencies will be calculated). This parameter will set the following variables (`carray` objects); for example, `alleleNum[1][2]` will be the number of allele `2` at locus `1`:

- `alleleNum[a]`, `subPop[sp]['alleleNum'][a]`
- `alleleFreq[a]`, `subPop[sp]['alleleFreq'][a]`.

alleleFreq_param A dictionary of parameters of `alleleFreq` statistics. Can be one or more items choosen from the following options: `numOfAlleles`, `alleleNum`, and `alleleFreq`.

association Association measures

association_param A dictionary of parameters of `association` statistics. Can be one or more items choosen from the following options: `ChiSq`, `ChiSq_P`, `UC_U`, and `CramerV`.

expHetero An array of loci at which the expected heterozygosities will be calculated (`expHetero=[loc1, loc2, ...]`). The expected heterozygosity is calculated by

$$h_{exp} = 1 - p_i^2.$$

The following variables will be set:

- `expHetero[loc]`, `subPop[sp]['expHetero'][loc]`.

expHetero_param A dictionary of parameters of `expHetero` statistics. Can be one or more items choosen from the following options: `subpop` and `midValues`.

genoFreq An array of loci at which all genotype frequencies will be calculated (`genoFreq=[loc1, loc2, ...]` where `loc1` etc. are loci where genotype frequencies will be calculated). All the genotypes in the population will be counted. You may use `hasPhase` to set if a/b and b/a are the same genotype. This parameter will set the following dictionary variables. Note that unlike list used for `alleleFreq` etc., the indexes `a`, `b` of `genoFreq[a][b]` are dictionary keys, so you will get a *KeyError* when you used a wrong key. Usually, `genoNum.setDefault(a,{})` is preferred.

- `genoNum[a][geno]` and `subPop[sp]['genoNum'][a][geno]`, the number of genotype `geno` at allele `a`. `geno` has the form `x-y`.
- `genoFreq[a][geno]` and `subPop[sp]['genoFreq'][a][geno]`, the frequency of genotype `geno` at allele `a`.

haploFreq A matrix of haplotypes (allele sequences on different loci) to count. For example, `haploFreq = [ [ 0,1,2 ], [1,2] ]` will count all haplotypes on loci 0,1 and 2; and all haplotypes on loci 1, 2. If only one haplotype is specified, the outer `[]` can be omitted. I.e., `haploFreq=[0,1]` is acceptable. The following dictionary variables will be set with keys `0-1-2` etc. For example, `haploNum['1-2']['5-6']` is the number of allele pair 5,6 (on loci 1 and 2 respectively) in the population.

- `haploNum[haplo]` and `subPop[sp]['haploNum'][haplo]`, the number of allele sequencies on loci `haplo`.
- `haploFreq[haplo]`, `subPop[sp]['haploFreq'][haplo]`, the frequency of allele sequencies on loci `haplo`.

hasPhase If a/b and b/a are the same genotype. Default to `False`.

heteroFreq An array of loci to calaulate observed heterozygosities and expected heterozygosities (`heteroFreq=[loc1, loc2, ...]`). This parameter will set the following variables (arrays of observed heterozygosities). Note that `heteroNum[loc][1]` is the number of heterozygote **1x**, $x \neq 1$. Numbers and frequencies (proportions) of heterozygotes are calculated for each allele. `HeteroNum[loc]` and `HeterFreq[loc]` are the overall heterozygosity number and frequency. I.e., the number/frequency of genotype **xy**, $x \neq y$. From this number, we can easily derive the number of homozygosity.

- `HeteroNum[loc]`, `subPop[sp]['HeteroNum'][loc]`, the overall heterozygote number
- `HeteroFreq[loc]`, `subPop[sp]['HeteroFreq'][loc]`, the overall heterozygote frequency
- `heteroNum[loc][allele]`, `subPop[sp]['heteroNum'][loc][allele]`
- `heteroFreq[loc][allele]`, `subPop[sp]['heteroFreq'][loc][allele]`

homoFreq An array of loci to calaulate observed homozygosities and expected homozygosities (`homoFreq=[loc1, loc2, ...]`). This parameter will calculate the numbers and frequencies of homozygotes **xx** and set the following variables:

- `homoNum[loc]`, `subPop[sp]['homoNum'][loc]`,
- `homoFreq[loc]`, `subPop[sp]['homoFreq'][loc]`.

midValues Whether or not post intermediate results. Default to `False`. For example, Fst will need to calculate allele frequencise. If `midValues` is set to `True`, allele frequencies will be posted as well. This will be helpful in debugging and sometimes in deriving statistics.

numOfAffected Whether or not count the numbers/proportions of affected and unaffected individuals. This parameter can set the following variables by user's specification:

- `numOfAffected`, `subPop[sp]['numOfAffected']` the number of affected individuals in the population/subpopulation
- `numOfUnaffected`, `subPop[sp]['numOfUnAffected']` the number of unaffected individuals in the population/subpopulation
- `propOfAffected`, `subPop[sp]['propOfAffected']` the proportion of affected individuals in the population/subpopulation
- `propOfUnaffected`, `subPop[sp]['propOfUnAffected']` the proportion of unaffected individuals in the population/subpopulation

numOfAffected_param A dictionary of parameters of `numOfAffected` statistics. Can be one or more items choosen from the following options: `numOfAffected`, `propOfAffected`, `numOfUnaffected`, `propOfUnaffected`.

numOfAlleles An array of loci at which the numbers of distinct alleles will be counted (`numOfAlleles=[loc1, loc2, ...]` where `loc1` etc. are absolute locus indexes). This is done through the calculation of allele frequencies. Therefore, allele frequencies will also be calculated if this statistics is requested. This parameter will set the following variables (`carray` objects of the numbers of alleles for *allloci*. Unrequested loci will have `0` distinct alleles.):

- `numOfAlleles`, `subPop[sp]['numOfAlleles']`, number of distinct alleles at each locus. (Calculated only at requested loci.)

numOfAlleles_param A dictionary of parameters of `numOfAlleles` statistics. Can be one or more items choosen from the following options: `numOfAffected`, `propOfAffected`, `numOfUnaffected`, `propOfUnaffected`.

numOfMale Whether or not count the numbers/proportions of males and females. This parameter can set the following variables by user's specification:

- `numOfMale`, `subPop[sp]['numOfMale']` the number of males in the population/subpopulation
- `numOfFemale`, `subPop[sp]['numOfFemale']` the number of females in the population/subpopulation.
- `propOfMale`, `subPop[sp]['propOfMale']` the proportion of males in the population/subpopulation
- `propOfFemale`, `subPop[sp]['propOfFemale']` the proportion of females in the population/-subpopulation

numOfMale_param A dictionary of parameters of `numOfMale` statistics. Can be one or more items choosen from the following options: `numOfMale`, `propOfMale`, `numOfFemale`, and `propOfFemale`.

popSize Whether or not calculate population sizes. This parameter will set the following variables:

- `numSubPop` the number of subpopulations
- `subPopSize` an array of subpopulation sizes. Not available for subpopulations.
- `popSize`, `subPop[sp]['popSize']` population/subpopulation size.

relGroups Calculate pairwise relatedness between groups. Can be in the form of either `[[1,2,3],[5,6,7],[8,9]]` or `[2,3,4]`. The first one specifies groups of individuals, while the second specifies subpopulations. By default, relatedness between subpopulations is calculated.

relLoci Loci on which relatedness values are calculated

relMethod Method used to calculate relatedness. Can be either `REL_Queller` or `REL_Lynch`. The relatedness values between two individuals, or two groups of individuals are calculated according to Queller & Goodnight (1989) (`method=REL_Queller`) and Lynch et al. (1999) (`method=REL_Lynch`). The results are pairwise relatedness values, in the form of a matrix. Original group or subpopulation numbers are discarded. `relatedness[grp1][grp2]` is the relatedness value between `grp1` and `grp2`. There is no subpopulation level relatedness values.

rel_param A dictionary of parameters of relatedness statistics. Can be one or more items choosen from the following options: `Fst`, `Fis`, `Fit`, `AvgFst`, `AvgFis`, and `AvgFit`.

Member Functions

`x.apply(pop)` Apply the `stat` operator

`x.clone()` Deep copy of a `stat` operator

2.11 Expression and Statements

2.11.1 Class `dumper`

Dump the content of a population.

Initialization

Dump population

```
dumper(alleleOnly=False, infoOnly=False, ancestralPops=False,
dispWidth=1, max=100, chrom=[], loci=[], subPop=[], indRange=[],
output=">", outputExpr="", stage=PostMating, begin=0, end=-1, step=1,
at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

alleleOnly Only display allele

ancestralPops Whether or not display ancestral populations, default to False

chrom Chromsoome(s) to display

dispWidth Width of allele display, default to 1

indRange Range(s) of individuals to display

infoOnly Only display info

loci Loci to display

max Max number of individuals to display, default to 100. This is to avoid careless dump of huge populations.

output Output file, default to standard output.

outputExpr And other parameters: refer to help(baseOperator.__init__)

subPop Only display subPop(s)

Member Functions

`x.alleleOnly()` Only show alleles (not structure, gene information?

`x.apply(pop)` Apply to one population. It does not check if the operator is activated.

`x.clone()` Deep copy of an operator

`x.infoOnly()` Only show info

`x.setAlleleOnly(alleleOnly)` SimuPOP::dumper::setAlleleOnly

`x.setInfoOnly(infoOnly)` SimuPOP::dumper::setInfoOnly

2.11.2 Class `savePopulation`

Save population to a file

Initialization

SimuPOP::savePopulation::savePopulation

```
savePopulation(output="", outputExpr="", format="bin", compress=True,
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply to one population. It does not check if the operator is activated.

`x.clone()` Deep copy of an operator

2.11.3 Class pyOutput

Output a given string.

Details

A common usage is pyOutpue('

', rep=REP_LAST)

Initialization

Create a pyOutput operator that output a given string.

```
pyOutput(str="", output=">", outputExpr="", stage=PostMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

str String to be outputted

Member Functions

x.apply(pop) Simply output some info

x.clone() Deep copy of an operator

x.setString(str) Set output string.

2.11.4 Class pyEval (Function form: PyEval)

Evaluate an expression

Details

Python expressions/statements will be executed when pyEval is applied to a population by using parameters expr/stmts. Statements can also been executed when pyEval is created and destroyed or before expr is executed. The corresponding parameters are preStmts, postStmts and stmts. For example, operator varPlotter uses this feature to initialize R plots and save plots to a file when finished.

Initialization

Evaluate expressions/statments in the local namespace of a replicate

```
pyEval(expr="", stmts="", preStmts="", postStmts="", exposePop=False,
name="", output=">", outputExpr="", stage=PostMating, begin=0,
end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

exposePop If true, expose current population as variable pop

expr The expression to be evaluated. Its result will be sent to output.

name Used to let pure Python operator to identify themselves

output Default to >. I.e., output to standard output.

postStmts The statement that will be executed when the operator is destroyed

preStmts The statement that will be executed when the operator is constructed

stmts The statement that will be executed before the expression

Member Functions

`x.apply(pop)` Apply the `pyEval` operator

`x.clone()` Deep copy of a `pyEval` operator

`x.name()` Return the name of an expression

The name of a `pyEval` operator is given by an optional parameter `name`. It can be used to identify this `pyEval` operator in debug output, or in the dryrun mode of simulator::evolve.

2.11.5 Class `pyExec` (Function form: `PyExec`)

Execute a Python statement

Details

This operator takes a list of statements and execute them. No value will be returned or outputted.

Initialization

Evaluate statments in the local replicate namespace, no return value

```
pyExec(stmts="", preStmts="", postStmts="", exposePop=False, name="",
output=">", outputExpr="", stage=PostMating, begin=0, end=-1, step=1,
at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

default To `>`. I.e., output to standard output.

exposePop If true, expose current population as variable `pop`

postStmts The statement that will be executed when the operator is destroyed

preStmts The statement that will be executed when the operator is constructed

stmts The statements (a single or multi-line string) that will be executed when this operator is applied.

Member Functions

`x.clone()` Deep copy of a `pyExec` operator

2.12 Tagging (used for pedigree tracking)

2.12.1 Class `tagger`

Base class of tagging individuals

Details

`TAGGER` is a during mating operator that tag individuals with various information. Potential usages are:

- recording parental information to track pedigree;
- tagging an individual/allele and monitor its spread in the population etc.

Initialization

Create a `tagger`, default to be always active but no output

```
tagger(begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

Member Functions

`x.clone()` Deep copy of a tagger

2.12.2 Class `inheritTagger`

Inherite tag from parents.

Details

This during-mating operator will copy the tag information from his/her parents. Depending on `mode` parameter, this tagger will obtain tag from his/her father (two tag fields), mother (two tag fields) or both (first tag field from both father and mother).An example may be tagging one or a few parents and see, at the last generation, how many offspring they have.

Initialization

Create an `inheritTagger`, default to be always active

```
inheritTagger(mode=TAG_Paternal, begin=0, end=-1, step=1,
at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=["paternal_tag",
"maternal_tag"])
```

Create a inheritTagger that inherit a tag from one or both parents. A tag is actually a information field whose value will be copied from parents to offspring. By default, paternal tag is copied to offspring's using the specified information field. If `mode=TAG_Both`, two tags will be copied from parents (info1 from father, and info2 from mother).

mode Can be one of `TAG_Paternal`, `TAG_Maternal`, and `TAG_Both`

Member Functions

`x.clone()` Deep copy of a `inheritTagger`

2.12.3 Class `parentsTagger`

Tagging according to parents' indexes

Details

This during-mating operator set
c tag(), currently a pair of numbers, of each individual with indexes of his/her parents in the parental population. This information will be used by pedigree-related operators like `affectedSibpairSample` to track the pedigree information. Since parental population will be discarded or stored after mating, and tagging information will be passed with individuals, mating/population change etc. will not interfere with this simple tagging system.

Initialization

Create a `parentsTagger`, default to be always active

```
parentsTagger(begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=["father_idx", "mother_idx"])
```

Member Functions

`x.clone()` Deep copy of a `parentsTagger`

Details

This tagger takes some information fields from both parents, pass to a Python function and set the individual field with the returned value.

This operator can be used to trace the inheritance of trait values.

Initialization

```
pyTagger(func=None, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

Creates a pyTagger that work on specified information fields.

func A Pyton function that returns a list to assign the information fields. e.g., if `fields=['A', 'B']`, the function will pass values of fields `'A'` and `'B'` of father, followed by mother if there is one, to this function. The returned value is assigned to fields `'A'` and `'B'` of the offspring. The returned value has to be a list even if only one field is given.

infoFields Information fields. The user should gurantee the existence of these fields.

Member Functions

`x.clone()` Deep copy of a `pyTagger`

2.13 Terminator

2.13.1 Class `terminator`

Base class of all terminators.

Details

These operators are used to see if an evolution is running as expected, and terminate the evolution if a certain condition fails.

Initialization

Create a terminator

```
terminator(message="", output=">", outputExpr="", stage=PostMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

message A message that will be displayed when the evolution is terminated.

Member Functions

`x.clone()` Deep copy of a terminator

2.13.2 Class `terminateIf`

Terminate according to a condition

Details

This operator terminates the evolution under certain conditions. For example, `terminateIf(condition='alleleFreq[0][1]<0.05', begin=100)` terminates the evolution if the allele frequency of allele `1` at locus `0` is less than 0.05. Of course, to make this opertor work, you will need to use a `stat` operator before it so that variable `alleleFreq` exists in the local namespace.

When the condition is true, a shared variable `var="terminate"` will be set to the current generation.

Initialization

Create a `terminateIf` terminator

```
terminateIf(condition="", message="", var="terminate", output="",
outputExpr="", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply the `terminateIf` terminator

`x.clone()` Deep copy of a `terminateIf` terminator

2.13.3 Class `continueIf`

Terminate according to a condition failure

Details

The same as `terminateIf` but continue if the condition is `True`.

Initialization

Create a `continueIf` terminator

```
continueIf(condition="", message="", var="terminate", output="",
outputExpr="", stage=PostMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply this operator

`x.clone()` Deep copy of a `continueIf` terminator

2.14 Conditional operator

2.14.1 Class `ifElse`

Conditional operator

Details

This operator accepts

- an expression that will be evaluated when this operator is applied;
- an operator that will be applied if the expression is `True` (default to null);
- an operator that will be applied if the expression is `False` (default to null).

When this operator is applied to a population, it will evaluate the expression and depending on its value, apply the supplied operator. Note that the `begin`, `at`, `step`, and `at` parameters of `ifOp` and `elseOp` will be ignored. For example, you can mimic the `at` parameter of an operator by `ifElse('rep in [2,5,9]' operator)`. The real use of this machanism is to monitor the population statistics and act accordingly.

Initialization

SimuPOP::ifElse::ifElse

```
ifElse(cond, ifOp=None, elseOp=None, output=">", outputExpr="",
stage=PostMating, begin=0, end=-1, step=1, at=[], rep=REP_ALL,
grp=GRP_ALL, infoFields=[])
```

cond Expression that will be treated as a bool variable

elseOp An operator that will be applied when `cond` is `False`

ifOp An operator that will be applied when `cond` is `True`

Member Functions

`x.apply(pop)` Apply the `ifElse` operator to one population

`x.clone()` Deep copy of an `ifElse` operator

Example

Example 2.14: Use of conditional operator

```
>>> simu = simulator(
...     population(size=1000, loci=[1]),
...     randomMating(), rep=4)
>>> simu.evolve(
...   preOps = [ initByValue([1,1])],
...   ops = [
...     # penetrance, additve penetrance
...     maPenetrance(locus=0, wildtype=[1], penetrance=[0,0.5,1]),
...     # count number of affected
...     stat(numOfAffected=True),
...     # introduce disease if no one is affected
...     ifElse(cond='numOfAffected==0',
...       ifOp=kamMutator(rate=0.01, maxAllele=2)),
...     ifElse(cond='numOfAffected==0',
...         ifOp=pyEval(r'"No affected at gen %d\n" % gen'))
...   ],
...   end=50
... )
No affected at gen 0
No affected at gen 0
No affected at gen 0
No affected at gen 0
```

```
No affected at gen 19
No affected at gen 22
No affected at gen 26
No affected at gen 32
True
>>>
```

2.15 Debug-related operators/functions

Standard `simuPOP` module can print out lots of debug information upon request. These are mostly used for internal debugging purposes but you can also use them when error happens. For example, the following code will crash simuPOP:

```
>>> population(1).individual(0).arrAllele()
```

It is not clear why this simple line will cause us trouble, instead of outputting the genotype of the only individual of this population. However, the reason is clear if you turn on debug information:

Example 2.15: Turn on/off debug information

```
>>> TurnOnDebug(DBG_ALL)
Debug code DBG_ALL is turned on. cf. listDebugCode(), turnOffDebug()
>>> population(1).individual(0).arrAlleles()
Constructor of Population is called
Population size 1
Destructor of Population is called
Segmentation fault (core dumped)
```

`population(1)` creates a temporary object that is destroyed right after the execution of the input. When Python tries to display the genotype, it will refer to an invalid location. The right way to do this is to create a persistent population object:

```
>>> pop = population(1)
>>> pop.individual(0).arrAllele()
```

If the output is overwhelming after you turn on all debug information, you can turn on certain part of the information by using the following functions:

- `ListDebugCode()` list all debug code.
- `turnOnDebug(), TurnOnDebug(code)` turn on debug codes.
- `turnOffDebug(), TurnOffDebug(code)` turn off debug codes.

`turnOnDebug()` and `turnOffDebug()` are operators and accept all operator parameters `begin, step` etc. Usually, you can use `turnOnDebug` to output more information about a potential bug before simuPOP starts to misbehave.

Another useful debug code is `DBG_PROFILE.` When turned on, it will display running time of each operator. This will give you a good sense of which operator runs slowly (or simply the order of operator execution if you are not sure). If most of the execution time is spent on a pure-Python operator, you may want to rewrite it in C++. Note that `DBG_PROFILE` is suitable for measuring individual operator performance. If you would like to measure the execution time of all operators in several generations, `ticToc` operator is better.

2.15.1 Class `turnOnDebug` (Function form: `TurnOnDebug`)

Set debug on

Details

Turn on debug. There are several ways to turn on debug information for non-optimized modules, namely

- set environment variable `SIMUDEBUG`
- use `simuOpt.setOptions(debug)` function, or
- use `TurnOnDebug` or `TurnOnDebugByName` function
- use this `turnOnDebug` operator

The advantage of using this operator is that you can turn on debug at given generations.

Initialization

SimuPOP::turnOnDebug::turnOnDebug

```
turnOnDebug(code, stage=PreMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply the `turnOnDebug` operator to one population

`x.clone()` Deep copy of a `turnOnDebug` operator

2.15.2 Class `turnOffDebug` (Function form: `TurnOffDebug`)

Set debug off

Details

Turn off debug.

Initialization

SimuPOP::turnOffDebug::turnOffDebug

```
turnOffDebug(code, stage=PreMating, begin=0, end=-1, step=1, at=[],
rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply the `turnOffDebug` operator to one population

`x.clone()` Deep copy of a `turnOffDebug` operator

2.16 Miscellaneous

2.16.1 Class `noneOp`

None operator

Initialization

```
noneOp(output=">", outputExpr="", stage=PostMating, begin=0, end=0,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

This operator does nothing.

Member Functions

`x.apply(pop)` Apply the `noneOp` operator to one population

`x.clone()` Deep copy of a `noneOp` operator

Example

Example 2.16: Use of noneOp operator

```
>>> # this may be set from command line option
>>> savePop = False
>>> # then, saveOp is defined accordingly
>>> if savePop:
...     saveOp = savePopulation(output='a.txt')
... else:
...     saveOp = noneOp()
...
>>> simu = simulator(population(10), randomMating())
>>> simu.step([saveOp])
True
>>>
```

2.16.2 Class `pause`

Pause a simulator

Details

This operator pauses the evolution of a simulator at given generations or at a key stroke, using `stopOnKeyStroke=True` option. Users can use `'q'` to stop an evolution. When a simulator is stopped, press any other key to resume the simulation or escape to a Python shell to examine the status of the simulation by press `'s'`.

There are two ways to use this operator, the first one is to pause the simulation at specified generations, using the usual operator parameters such as `at`. Another way is to pause a simulation with any key stroke, using the `stopOnKeyStroke` parameter. This feature is useful for a presentation or an interactive simulation. When `'s'` is pressed, this operator expose the current population to the main Python dictionary as variable `pop` and enter an interactive Python session. The way current population is exposed can be controlled by parameter `exposePop` and `popName`. This feature is useful when you want to examine the properties of a population during evolution.

Initialization

Stop a simulation. Press `'q'` to exit or any other key to continue.

```
pause(prompt=True, stopOnKeyStroke=False, exposePop=True,
popName="pop", output=">", outputExpr="", stage=PostMating, begin=0,
end=-1, step=1, at=[], rep=REP_LAST, grp=GRP_ALL, infoFields=[])
```

exposePop Whether or not expose `pop` to user namespace, only useful when user choose `'s'` at pause. Default to `True`.

popName By which name the population is exposed. Default to `pop`.

prompt If `True` (default), print prompt message

stopOnKeyStroke If `True`, stop only when a key was pressed

Member Functions

`x.apply(pop)` Apply the `pause` operator to one population

`x.clone()` Deep copy of a `pause` operator

2.16.3 Class `ticToc` (Function form: `TicToc`)

Timer operator

Details

This operator, when called, output the difference between current and the last called clock time. This can be used to estimate execution time of each generation. Similar information can also be obtained from `turnOnDebug(DBG_PROFILE)`, but this operator has the advantage of measuring the duration between several generations by setting `step` parameter.

Initialization

Create a timer

```
ticToc(output=">", outputExpr="", stage=PreMating, begin=0, end=-1,
step=1, at=[], rep=REP_ALL, grp=GRP_ALL, infoFields=[])
```

Member Functions

`x.apply(pop)` Apply the `ticToc` operator to one population

`x.clone()` Deep copy of a `ticToc` operator

2.16.4 Class `setAncestralDepth`

Set ancestral depth

Details

This operator set the number of ancestral generations to keep in a population. It is usually called like `setAncestral(at=[-2])` to start recording ancestral generations to a population at the end of the evolution. This is useful when constructing pedigree trees from a population.

Initialization

Create a `setAncestralDepth` operator

```
setAncestralDepth(depth, output=">", outputExpr="", stage=PreMating,
begin=0, end=-1, step=1, at=[], rep=REP_ALL, grp=GRP_ALL,
infoFields=[])
```

Member Functions

`x.apply(pop)` Apply the `setAncestralDepth` operator to one population

`x.clone()` Deep copy of a `setAncestralDepth` operator

CHAPTER

THREE

Global and Python Utility functions

3.1 Global functions

`AlleleType()`

Return the allele type of the current module. Can be `binary`, `short`, or `long`.

`Limits()`

Print out system limits

`ListAllRNG()`

List the name of all available random number generators

`ListDebugCode()`

List all debug codes

`LoadPopulation(file, format="auto")`

Load a population from a file. The file format is by default determined by file extension (`format="auto"`). Otherwise, `format` can be one of `txt`, `bin`, or `xml`.

`LoadSimulator(file, mate, format="auto")`

Load a simulator from a file with the specified mating scheme. The file format is by default determined by file extension (`format="auto"`). Otherwise, `format` can be one of `txt`, `bin`, or `xml`.

`MaxAllele()`

Return 1, 2^8-1, $2^{16}-1$ for binary, short, or long allele modules, respectively

`MergePopulations(pops, newSubPopSizes=[], keepAncestralPops=-1)`

Merge several populations with the same genotypic structure and create a new population

`MergePopulationsByLoci(pops, newNumLoci=[], newLociPos=[], byChromosome=False)`

Merge several populations of the same size by loci and create a new population

`ModuleCompiler()`

Return the compiler used to compile this simuPOP module

`ModuleDate()`

Return the date when this simuPOP module is compiled

`ModulePlatForm()`

Return the platform on which this simuPOP module is compiled

ModulePyVersion()

Return the Python version this simuPOP module is compiled for

Optimized()

Return `True` if this simuPOP module is optimized

SetRNG(rng="", seed=0)

Set random number generator. If `seed=0` (default), a random seed will be given. If `rng=""`, seed will be set to the current random number generator.

TurnOffDebug(code=DBG_ALL)

Turn off debug information. Default to turn off all debug codes. Only available in non-optimized modules. Do not mix this function with `turnOffDebug()`, which creates an operator

TurnOnDebug(code=DBG_ALL)

Set debug codes. Default to turn on all debug codes. Only available in non-optimized modules. Do not mix this function with `turnOnDebug()`, which creates an operator

rng()

Return the currently used random number generator

simuRev()

Return the revision number of this simuPOP module. Can be used to test if a feature is available.

simuVer()

Return the version of this simuPOP module

3.2 Utility Modules

Several utility modules are distributed with simuPOP. They provide important functions and extensions to simuPOP and serve as good examples on how simuPOP can be used.

Compared to simuPOP kernel functions, these utility functions are less tested, and are subject to more frequent changes. Please report to simuPOP mailing list if any function stops working.

3.2.1 Module `simuOpt`

Module `simuOpt` can be used to control which simuPOP module to load, and how it is loaded using function `setOptions` . It also provides a simple way to set simulation options, from user input, command line, configuration file or a parameter dialog. All you need to do is to define an option description list that lists all parameters in a given format, and call the `getParam` function. This module, if loaded, pre-process the command line options. More specifically, it checks command line option:

-c configfile read from a configuration file

--config configgile the same as -c

--optimized load optimized modules, unless setOption explicitly use non-optimizedmodules.

-q Do not display banner information when simuPOP is loaded

--quiet the same as -q

--useTkinter force the use of Tcl/Tk dialog even when wxPython is available. By default, wxPython is used whenever possible.

--noDialog do not use option dialog. If the options can not be obtained fromcommand line or configuraiton file, users will be asked to input them interactively.

Because these options are reserved, you can not use them in your simuPOP script.

Module Functions

getParam (options=[], doc=", details=", noDialog=False, checkUnprocessedArgs=True, verbose=False, nCol=1)

Get parameters from either

- a Tcl/Tk based, or wxPython based parameter dialog (wxPython is used if it is available)
- command line argument
- configuration file specified by `-c file` (`-config` file), or
- prompt for user input

The option description list consists of dictionaries with some predefined keys. Each dictionary defines an option. Each option description item can have the following keys:

arg short command line option name. `'h'` checks the presence of argument `-h` . If an argument is expected, add a comma to the option name. For example, `'p:'` matches command line option `-p=100` or `-p 100` .

longarg long command line option name. `'help'` checks the presence of argument `'-help'` . `'mu='` matches command line option `-mu=0.001` or `-mu 0.001` .

label The label of the input field in a parameter dialog, and as the prompt foruser input.

default default value for this parameter. It is used to as the default valuein the parameter dialog, and as the option value when a user presses `'Enter'` directly during interactive parameter input.

useDefault use default value without asking, if the value can not be determinedfrom GUI, command line option or config file. This is useful for options that rarely need to be changed. Setting them to useDfault allows shorter command lines, and easy user input.

description a long description of this parameter, will be put into the usage information, which will be displayed with (`-h` , `-help` command line option, or help button in parameter dialog).

allowedTypes acceptable types of this option. If `allowedTypes` is `types.ListType` or `types.TupleType` and the user's input is a scalar, the input will be converted to a list automatically. If the conversion can not be done, this option will not be accepted.

validate a function to validate the parameter. You can define your own functions or use the ones defined in this module.

chooseOneOf if specified, `simuOpt` will choose one from a list of values using a listbox (Tk) or a combo box (wxPython) .

chooseFrom if specified, `simuOpt` will choose one or more items from a list ofvalues using a listbox (tk) or a combo box (wxPython).

separator if specified, a blue label will be used to separate groups of parameters.

jump it is used to skip some parameters when doing the interactive user input. For example, `getParam` will skip the rest of the parameters if `-h` is specified if parameter `-h` has item `'jump':-1` which means jumping to the end. Another situation of using this value is when you have a hierarchical parameter set. For example, if mutation is on, specify mutation rate, otherwise proceed.

jumpIfFalse The same as jump but jump if current parameter is `False` .

This function will first check command line argument. If the argument is available, use its value. Otherwise check if a config file is specified. If so, get the value from the config file. If both failed, prompt user to input a value. All input will be checked against types, if exists, an array of allowed types. Parameters of this function are:

options a list of option description dictionaries

doc short description put to the top of parameter dialog

details module help. Usually set to `__doc__` .

noDialog do not use a parameter dialog, used in batch mode. Default to False.

checkUnprocessedArgs check args, avoid misspelling of arg name

verbose whether or not print detailed info

nCol number of columns in the parameter dialog.

printConfig (opt, param, out=<open file '<stdout>', mode 'w' at 0x2a955940a0>)

Print configuration.

opt option description list

param parameters returned from `getParam()`

out output

requireRevision (rev)

Compare the revision of this simuPOP module with given revision. Raise an exception if current module is out of date.

saveConfig (opt, file, param)

Write a configuration file. This file can be later read with command line option `-c` or –config `.`

opt the option description list

file output file

param parameters returned from `getParam`

setOptions (optimized=None, mpi=None, chromMap=[], alleleType=None, quiet=None, debug=[])

set options before simuPOP is loaded to control which simuPOP module to load, and how the module should be loaded.

optimized whether or not load optimized version of a module. If not set,environmental variable SIMUOPTIMIZED, and commandline option –optimized will be used if available. If nothing is defined, standard version will be used.

alleleType 'binary', 'short', or 'long'. 'standard' can be used as 'short'for backward compatibility. If not set, environmental variable SIMUALLELETYPE will be used if available. if it is not defined, the short allele version will be used.

quiet If True, supress banner information when simuPOP is loaded.

debug a list of debug code (or string). If not set, environmental variableSIMUDEBUG will be used if available.

mpi currently unused

chromMap currently unused

usage (options, before=")

Print usage information from the option description list. Used with `-h` (or `-help`) option, and in the parameter input dialog.

options option description list.

before optional information

valueAnd (t1, t2)

Return a function that returns true if passed option passes validator t1 and t2

valueBetween (a, b)

Return a function that returns true if passed option is between value a and b (a and b included)

valueEqual (a)

Return a function that returns true if passed option equals a

valueGE (a)

Return a function that returns true if passed option is greater than or equal to a

valueGT (a)

Return a function that returns true if passed option is greater than a

valueIsList ()

Return a function that returns true if passed option is a list (or tuple)

valueIsNum ()

Return a function that returns true if passed option is a number (int, long or float)

valueLE (a)

Return a function that returns true if passed option is less than or equal to a

valueLT (a)

Return a function that returns true if passed option is less than a

valueListOf (t)

Return a function that returns true if passed option val is a list of type t. If t is a function (validator), check if all v in val pass t(v)

valueNot (t)

Return a function that returns true if passed option does not passes validator t

valueNotEqual (a)

Return a function that returns true if passed option does not equal a

valueOneOf (t)

Return a function that returns true if passed option is one of the values list in t

valueOr (t1, t2)

Return a function that returns true if passed option passes validator t1 or t2

valueTrueFalse ()

Return a function that returns true if passed option is True or False

valueValidDir ()

Return a function that returns true if passed option val if a valid directory

valueValidFile ()

Return a function that returns true if passed option val if a valid file

3.2.2 Module `simuUtil`

This module provides some commonly used operators and format conversion utilities.

Module Functions

`CaseControl_ChiSq (pop, sampleSize, penetrance=None)`

Draw affected sibpair sample from pop, run TDT using GENEHUNTER

`pop` simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

`pene` penetrance function, if not given (None), existing affectionstatus will be used.

`sampleSize` total sample size N. N/4 is the number of families to ascertain.

`keep_temp` if True, do not remove sample data. Default to False.

`ChiSq_test (pop)`

perform case control test Parameters;

`pop` loaded population, or population file in simuPOP format. This function assumes that pop has two subpopulations, cases and controls, and have 0 as wildtype and 1 as disease allele. pop can also be an loaded population object.

`Return value` A list of p-value at each locus.

`Note` this function requires rpy module.

`CollectValue (pop, gen, expr, name)`

data collector

`ConstSize (size, split=0, numSubPop=1, bottleneckGen=-1, bottleneckSize=0)`

The population size is constant, but will split into numSubPop subpopulations at generation split

`ExponentialExpansion (initSize, endSize, end, burnin=0, split=0, numSubPop=1, bottleneckGen=-1, bottleneckSize=0)`

Exponentially expand population size from intiSize to endSize after burnin, split the population at generation split.

`FreqTrajectoryMultiStochWithSubPop (curGen, numLoci, freq, NtFunc, minMutAge, maxMutAge, fitness=[], mode='uneven', ploidy=2, restartIfFail=True, fitnessFunc=None)`

Simulate frequency trajectory with subpopulation structure, migration is currently ignored. The essential part of this script is to simulate the trajectory of each subpopulation independently by calling FreqTrajectoryMultiStoch with properly wrapped NtFunc function. If mode = 'even' (default) When freq is the same length of the number of loci. The allele frequency at the last generation will be multi-nomially distributed. If freq for each subpop is specified in the order of loc1-sp1, loc1-sp2, .. loc2-sp1, ... This freq will be used directly. If mode = 'uneven'. The number of disease alleles will be proportional to the interval lengths of 0 x x x 1 while x are uniform [0,1]. The distribution of interval lengths, are roughly exponential (conditional on overall length 1). ' If mode = 'none', subpop will be ignored. This script assume a single-split model of NtFunc

`InstantExpansion (initSize, endSize, end, burnin=0, split=0, numSubPop=1, bottleneckGen=-1, bottleneckSize=0)`

Instaneously expand population size from intiSize to endSize after burnin, split the population at generation split.

LOD_gh (file, gh='gh')

Analyze data using the linkage method of genehunter. Note that this function may not work under platforms other than linux, and may not work with your version of genehunter. As a matter of fact, it is almost unrelated to simuPOP and is provided only as an example how to use python to analyze data.

Parameters

file file to analyze. This function will look for file.dat and file.pre in linkage format.

loci a list of loci at which p-value will be returned. If the list is empty,all p-values are returned.

gh name (or full path) of genehunter executable. Default to 'gh'

Return value A list (for each chromosome) of list (for each locus) of p-values.

LOD_merlin (file, merlin='merlin')

run multi-point non-parametric linkage analysis using merlin

LargePeds_Reg_merlin (pop, sampleSize, qtrait=None, infoField='qtrait', merlin='merlin-regress', keep_temp=False)

Draw affected sibpair sample from pop, run TDT using GENEHUNTER

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

qtrait a function to calculate quantitative trait

infoField information field to store quantitative trait. Default to 'qtrait'

sampleSize total sample size N. N/4 is the number of families to ascertain.

merlin executable name of merlin, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

LargePeds_VC_merlin (pop, sampleSize, qtrait=None, infoField='qtrait', merlin='merlin', keep_temp=False)

Draw affected sibpair sample from pop, run TDT using GENEHUNTER

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

qtrait a function to calculate quantitative trait

infoField information field to store quantitative trait. Default to 'qtrait'

sampleSize total sample size N. N/4 is the number of families to ascertain.

merlin executable name of merlin, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

LinearExpansion (initSize, endSize, end, burnin=0, split=0, numSubPop=1, bottleneckGen=-1, bottleneckSize=0)

Linearly expand population size from intiSize to endSize after burnin, split the population at generation split.

ListVars (var, level=-1, name="", subPop=True, useWxPython=True)

list a variable in tree format, either in text format or in a wxPython window.

var any variable to be viewed. Can be a dw object returnedby dvars() function

level level of display.

name only view certain variable

subPop whether or not display info in subPop

useWxPython if True, use terminal output even if wxPython is available.

LoadFstat (file, loci=[])

load population from fstat file 'file' # since fstat does not have chromosome structure # an additional parameter can be given

LoadGCData (file, loci=[])

read GC data file in http://wpicr.wpic.pitt.edu/WPICCompGen/genomic_control/genomic_control.htm

MigrIslandRates (r, n)

migration rate matrix x m/(n-1) m/(n-1) .... m/(n-1) x .... m/(n-1) m/(n-1) x where x = 1-m

MigrSteppingStoneRates (r, n, circular=False)

migration rate matrix, circular step stone model (X=1-m) X m/2 m/2 m/2 X m/2 0 0 m/2 x m/2 .....0 ... m/2 0 .... m/2 X or non-circular X m/2 m/2 m/2 X m/2 0 0 m/2 X m/2 .....0 ... ... m X

QtraitSibs_Reg_merlin (pop, sampleSize, qtrait=None, infoField='qtrait', merlin='merlin-regress', keep_temp=False)

Draw affected sibpair sample from pop, run TDT using GENEHUNTER

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

qtrait a function to calculate quantitative trait

infoField information field to store quantitative trait. Default to 'qtrait'

sampleSize total sample size N. N/4 is the number of families to ascertain.

merlin executable name of merlin, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

QtraitSibs_VC_merlin (pop, sampleSize, qtrait=None, infoField='qtrait', merlin='merlin', keep_temp=False)

Draw affected sibpair sample from pop, run TDT using GENEHUNTER

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

qtrait a function to calculate quantitative trait

infoField information field to store quantitative trait. Default to 'qtrait'

sampleSize total sample size N. N/4 is the number of families to ascertain.

merlin executable name of merlin, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

Regression_merlin (file, merlin='merlin-regress')

run merlin regression method

SaveCSV (pop, output="", outputExpr="", fields=['sex', 'affection'], loci=[], combine=None, shift=1, **kwargs)

save file in CSV format

fileds information fields, 'sex' and 'affection' are special fields that is treated differently.

genotype list of loci to output, default to all.

combine how to combine the markers. Default to None.A function can be specified, that takes the form

def func(markers) return markers[0]+markers[1]

shift since alleles in simuPOP is 0-based, shift=1 is usually needed to output alleles starting from allele 1. This parameter is ignored if combine is used.

SaveFstat (pop, output=”, outputExpr=”, maxAllele=0, loci=[], shift=1, combine=None)

save file in FSTAT format

SaveLinkage (pop, output=”, outputExpr=”, loci=[], shift=1, combine=None, fields=[], recombination=1.0000000000000001e-05, penetrance=[0, 0.25, 0.5], affectionCode=[’1’, ’2’], pre=True, daf=0.001)

save population in Linkage format. Currently only support affected sibpairs sampled with affectedSibpairSample operator.

pop population to be saved. Must have ancestralDepth 1.paired individuals are sibs. Parental population are corresponding parents. If pop is a filename, it will be loaded.

output output.dat and output.ped will be the data and pedigree file.You may need to rename them to be analyzed by LINKAGE. This allows saving multiple files.

outputExpr expression version of output.

affectionCode default to ’1’

pre True. pedigree format to be fed to makeped. Non-pre format it is likely to be wrong now for non-sibpair families.

Note the first child is always the proband.

SaveMerlinDatFile (pop, output=”, outputExpr=”, loci=[], fields=[], outputAffection=False)

Output a .dat file readable by merlin

SaveMerlinMapFile (pop, output=”, outputExpr=”, loci=[])

Output a .map file readable by merlin

SaveMerlinPedFile (pop, output=”, outputExpr=”, loci=[], fields=[], header=False, outputAffection=False, affectionCode=[’U’, ’A’], combine=None, shift=1, **kwargs)

Output a .ped file readable by merlin

SaveQTDT (pop, output=”, outputExpr=”, loci=[], header=False, affectionCode=[’U’, ’A’], fields=[], combine=None, shift=1, **kwargs)

save population in Merlin/QTDT format. The population must have pedindex, father_idx and mother_idx information fields.

pop population to be saved. If pop is a filename, it will be loaded.

output base filename.

outputExpr expression for base filename, will be evaluated in pop’slocal namespace.

affectionCode code for unaffected and affected. ’1’, ’2’ are default,but ’U’, and ’A’ or others can be specified.

loci loci to output

header whether or not put head line in the ped file.

fields information fields to output

combine an optional function to combine two alleles of a diploid individual.

shift if combine is not given, output two alleles directly, addingthis value (default to 1).

SaveSolarFrqFile (pop, output=”, outputExpr=”, loci=[], calcFreq=True)

Output a frequency file, in a format readable by solar

calcFreq whether or not calculate allele frequency

Sibpair_LOD_gh (pop, sampleSize, penetrance=None, recRate=None, daf=None, gh='gh', keep_temp=False)

Draw affected sibpair sample from pop, run Linkage analysis using GENEHUNTER

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

pene penetrance function, if not given (None), existing affectionstatus will be used.

sampleSize total sample size N. N/4 is the number of families to ascertain.

recRate recombination rate, used in the Linkage file. If not given,pop.dvars().recRate[0] will be used. If there is no such variable, 0.0001 is used.

daf disease allele frequency. This is needed for the linkage formatbut I am not sure if it is used by TDT.

gh executable name of genehunter, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

Sibpair_LOD_merlin (pop, sampleSize, penetrance=None, merlin='merlin', keep_temp=False)

Draw affected sibpair sample from pop, run multi-point linkage analysis using merlin

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

pene penetrance function, if not given (None), existing affectionstatus will be used.

sampleSize total sample size N. N/4 is the number of families to ascertain.

merlin executable name of merlin, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

Sibpair_TDT_gh (pop, sampleSize, penetrance=None, recRate=None, daf=None, gh='gh', keep_temp=False)

Draw affected sibpair sample from pop, run TDT using GENEHUNTER

pop simuPOP population. It can be a string if path to a file is given.This population must 1. have at least one ancestral generation (parental generation) 2. have a variable DSL (pop.dvars().DSL) indicating the Disease susceptibility loci. These DSL will be removed from the samples. 3. has only binary alleles

pene penetrance function, if not given (None), existing affectionstatus will be used.

sampleSize total sample size N. N/4 is the number of families to ascertain.

recRate recombination rate, used in the Linkage file. If not given,pop.dvars().recRate[0] will be used. If there is no such variable, 0.0001 is used.

daf disease allele frequency. This is needed for the linkage formatbut I am not sure if it is used by TDT.

gh executable name of genehunter, full path name can be given.

keep_temp if True, do not remove sample data. Default to False.

TDT_gh (file, gh='gh')

Analyze data using genehunter/TDT. Note that this function may not work under platforms other than linux, and may not work with your version of genehunter. As a matter of fact, it is almost unrelated to simuPOP and is provided only as an example how to use python to analyze data.

Parameters

file file to analyze. This function will look for file.dat and file.pre in linkage format.

loci a list of loci at which p-value will be returned. If the list is empty,all p-values are returned.

gh name (or full path) of genehunter executable. Default to 'gh'

Return value A list (for each chromosome) of list (for each locus) of p-values.

VC_merlin (file, merlin='merlin')

run variance component method

file file.ped, file.dat, file.map and file,mdl are expected.file can contain directory name.

collector (name, expr, **kwargs)

wrapper

dataAggregator (self, maxRecord=0, recordSize=0)

collect variables so that plotters can plot them all at once You can of course put it in other uses

Usage a = dataAggregator(maxRecord=0, recordSize=0)

maxRecord if more data is pushed, the old ones are discarded

recordSize size of recorda.push(gen, data, idx=-1)

gen generation number

data one record (will set recordSize if the first time), or

idx if idx!=-1, set data at idx.a.clear() a.range() # return min, max of all data a.data[i] # column i of the data a.gen # a.ready() # if all column has the same length, so data is ready

Internal data storage self.gen [....] self.data column1 [.....] column2 [.....] ..... each record is pushed at the end of

Initialization

maxRecord maxRecorddow size. I.e., maximum generations of data to keep

endl (output='>', outputExpr='', **kwargs)

getGenotype (pop, atLoci=[], subPop=[], indRange=[], atPloidy=[])

Obtain genotype as specified by parameters

atLoci subset of loci, default to all

subPop subset of subpopulations, default ao all

indRange individual ranges This is mostly used for testing purposes because the returned array can be large for large populations.

saveFstat (output='', outputExpr='', **kwargs)

operator version of the function SaveFstat

saveLinkage (output='', outputExpr='', **kwargs)

An operator to save population in linkage format

tab (output='>', outputExpr='', **kwargs)

operator tab (I can use operator output # but the name conflicts with parameter name # and I would not want to go through the trouble # of a walkaround (like aliasing output)

testDemoFunc (end, func)

for internal use only

trajFunc (endingGen, traj)

return freq at each generation from a simulated trajctories.

3.2.3 Module simuRPy

This module helps the use of rpy package with simuPOP. It defines an operator varPlotter that can be used to plot population expressions when rpy is installed.

Module Functions

rmatrix (mat)

Convert a Python 2d list to r matrix format that can be passed to functions like image directly.

varPlotter (self, expr, history=True, varDim=1, numRep=1, win=0, ylim=[0, 0], update=1, title='', xlab='generation', ylab='', axes=True, lty=[], col=[], mfrow=[1, 1], separate=False, byRep=False, byVal=False, plotType='plot', level=20, saveAs='', leaveOpen=True, dev='', width=0, height=0, *args, **kwargs)

Plotting with history plot a number in the form of a variable or expression, use

```
>>> varPlotter(var='expr')
```

plot a vector in the same window and there is only one replicate in the simulator, use

```
>>> varPlotter(var='expr', varDim=len)
```

where len is the dimension of your variable or expression. Each line in the figure represents the history of an item in the array. plot a vector in the same window and there are several replicates, use varPlotter(var='expr', varDim=len, numRep=nr, byRep=1) varPlotter will try to use an appropriate layout for your subplots (for example, use 3x4 if numRep=10). You can also specify parameter mfrow to change the layout. if you would like to plot each item of your array variables in a subplot, use varPlotter(var='expr', varDim=len, byVal=1) or in case of a single replicate varPlotter(var='expr', varDim=len, byVal=1, numRep=nr) There will be numRep lines in each subplot. Plotting without history use option history=False. Parameters byVal, varDim etc. will be ignored. Other options are

title, xtitle, ytitle title of your figure(s). title is default to your expression, xtitle is defaulted to generation.

win window of generations. I.e., how many generations to keep in a figure. This is useful when you want to keep track of only recent changes.

update update figure after update generations. This is used when you do not want to update the figure at every generation.

saveAs save figures in files saveAs#gen.eps. For example, if saveAs='demo', you will get files demo1.eps, demo2.eps etc.

separate plot data lines in separate panels.

image use R image function to plot image, instead of lines.

level level of image colors (default to 20).

leaveOpen whether or not leave the plot open when plotting is done. Default to True. Initialization

create a varplotter instance

3.2.4 Module `hapMapUtil`

Utility functions to manipulate HapMap data. These functions are provided as samples on how to load and evolve the HapMap dataset. They tend to change frequently so do not call these functions directly. It is recommended that you copy these function to your script when you need to use them.

Module Functions

`evolveHapMap (pop, endingSize, endGen, migr=<simuPOP_std.noneOp; proxy of <Swig Object of type 'simuPOP::noneOp *' at 0x43f0630> >, expand='exponential', mergeAt=10000, initMultiple=1, recIntensity=0.01, mutRate=9.9999999999999995e-08, step=10, keepParents=False, numOffspring=1)`

Evolve and expand the hapmap population

`gen` total evolution generation

`initMultiple` copy each individual initMultiple times, to avoidrapid loss of genotype variation when population size is small.

`endingSize` ending poplation size

`expand` expanding method, can be linear or exponential

`mergeAt` when to merge population?

`endGen` endingGeneration

`recIntensity` recombination intensity

`mutRate` mutation rate

`step` step at which to display statistics

`keepParents` whether or not keep parental generations

`numOffspring` number of offspring at the last generation

`migr` a migrator to be used.

`getMarkersFromName (hapmap_dir, names, chroms=[])`

Get population from marker names. This function returns a tuple with a population with found markers and names of markers that can not be located in the HapMap data. The returned population has three subpopulations, corresponding to CEU, YRI and JPT+CHB hapmap populations.

`hapmap_dir` where hapmap data in simuPOP format is stored. The filesshould have been prepared by scripts/loadHapMap.py.

`names` names of markers

`chroms` a list of chromosomes to look in. If empty, all 22 autosomeswill be tried.

`getMarkersFromRange (hapmap_dir, chrom, startPos, endPos, maxNum, minAF, minDist)`

Get a population with markers from given range

`hapmap_dir` where hapmap data in simuPOP format is stored. The filesshould have been prepared by scripts/loadHapMap.py.

`chrom` chromosome number (1-based index)

`startPos` starting position (in cM)

`endPos` ending position (in cM)

`maxNum` maximum number of markers to get

minAF minimal minor allele frequency

minDist minimal distance between two adjacent markers, in cM

sample1DSL (pop, DSL, DA, pene, name, sampleSize)

Sample from the final population, using a single lcous penetrance model.

DSL disease locus

DA disease allele

pene penetrance

name name of directory to save (it must exist)

sampleSize sample size, in this case, sampleSize/4 is the number of families

sample2DSL (pop, DSL, pene, name, size)

Sample from the final population, using a two locus penetrance model

DSL disease loci (two locus)

pene penetrance value, assuming a two-locus model

name name to save sample

size sample size

CHAPTER

FOUR

Extending simuPOP

simuPOP can be extended easily using Python programming language. Because almost all data are exposed to the Python interface, your ability of extending simuPOP is *unlimited*. However, because Python is slower than C++ and the exchange of data between internal C++ data structure and Python interface may be costly, it is not recommended to write frequently used operators in Python. Appropriate pure Python operators are visualizers, statistics calculators, file outputers etc.

To write simuPOP extension, you will have to know more about data structures and member functions of population. Note that for efficiency and implementation reasons, many of the following functions do not provide keyword parameters.

4.1 Genotypic structure

The genotypes of an individual are organized as a single array. For example, if you have an diploid individual with two chromosomes, having 2 and 3 loci respectively. The genotypes should be in the order of

```
0-0-0, 1-0-0, 0-1-0, 1-1-0, 2-1-0, 0-0-1, 1-0-1, 0-1-1, 1-1-1, 2-1-1,
```

where X-X-X are locus-chromosome-ploidy indices. An important consequence of this arrangement is that 'locus location' + 'the total number of loci' is the location of the locus on the other set of chromosomes.

Several functions are provided to retrieve genotypic information:

- `ploidy()`, ploidy
- `numChrom()`, the number of chromosomes
- `numLoci(chrom)`, the number of loci on chromosome `chrom`
- `totNumLoci()`, the total number of loci
- `genoSize()`, the size of genotype. Equals to `totNumLoci()*ploidy()`.
- `alleleName()`, allele name given by parameter `alleleNames`. Otherwise the allele number is returned.
- `locusPos(loc)`, the locus position on chromosome (Distance to the beginning of chromosome)
- `arrlociPos()`, returns an `carray` of the locus distances.

The last function is very interesting. It actually returns the reference of the internal locus distance array. If you change the values of the returned array, the internal locus distance will be changed! All functions with this property will be named `arrFunc()`.

The following example shows how to change the locus distance through this function.

Example 4.1: geno stru

```
>>> pop = population(1, loci=[2,3,4])
>>> print pop.numLoci(1)
3
>>> print pop.locusPos(2)
1.0
>>> dis = pop.arrLociPos()
>>> print dis
[1.0, 2.0, 1.0, 2.0, 3.0, 1.0, 2.0, 3.0, 4.0]
>>> dis[2] = 0.5
>>> print pop.locusPos(2)
0.5
>>> print pop.arrLociPos()
[1.0, 2.0, 0.5, 2.0, 3.0, 1.0, 2.0, 3.0, 4.0]
>>>
```

4.2 Accessing genotype and other info

Genotype of an individual can be retrieved through the following functions:

- `ind.allele(index, p=0)`,
- `ind.setAllele(value, index, p=0)`,
- `ind.arrGenotype(p=0, ch=0)`,

where `p` means ploidy. I.e., the index of the copy of chromosomes. `ch` means chromosome. For example

```
pop.individual(1).arrGenotype(1, 2)
```

returns an array of alleles on the third chromosome of the second copy of chromosomes, of the second individual in the population `pop`.

Example 4.2: genotype

```
>>> InitByFreq(pop, [.2,.8])
>>> Dump(pop, alleleOnly=1)
individual info:
sub population 0:
   0: FU   1  0   1  1  0   1  1  1  1 |   0  1   1  1  1   1  1  1  0
End of individual info.


No ancenstral population recorded.
>>> ind = pop.individual(0)
>>> print ind.allele(1,1)
1
>>> ind.setAllele(3,1,1)
>>> Dump(pop, alleleOnly=1)
individual info:
sub population 0:
   0: FU   1  0   1  1  0   1  1  1  1 |   0  3   1  1  1   1  1  1  0
End of individual info.
```

```
No ancenstral population recorded.
>>> a = ind.arrGenotype()
>>> print a
[1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 3, 1, 1, 1, 1, 1, 1, 0]
>>> a = ind.arrGenotype(1)
>>> print a
[0, 3, 1, 1, 1, 1, 1, 1, 0]
>>> a = ind.arrGenotype(1,2)
>>> print a
[1, 1, 1, 0]
>>> a[2]=4
>>> # the allele on the third chromosome has been changed
>>> Dump(pop, alleleOnly=1)
individual info:
sub population 0:
   0: FU   1  0   1  1  0   1  1  1  1 |   0  3   1  1  1   1  1  4  0
End of individual info.


No ancenstral population recorded.
>>>
```

Sex, affected status can be accessed through `sex, setSex, affected, setAffected` functions.

Example 4.3: genotype

```
>>> print ind.sex()
2
>>> print ind.sexChar()
F
>>> ind.setSex(Female)
>>> ind.setAffected(True)
>>> print ind.tag()
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
AttributeError: 'individual' object has no attribute 'tag'
>>> ind.setTag([1,2])
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
AttributeError: 'individual' object has no attribute 'setTag'
>>> Dump(pop)
Ploidy:                2
Number of chrom:       3
Number of loci:        2 3 4
Maximum allele state:  65535
Loci positions:
               1 2
               0.5 2 3
               1 2 3 4
Loci names:
               loc1-1 loc1-2
               loc2-1 loc2-2 loc2-3
```

```
               loc3-1 loc3-2 loc3-3 loc3-4
population size:        1
Number of subPop:       1
Subpop sizes:           1
Number of ancestral populations:        0
individual info:
sub population 0:
   0: FA   1  0   1  1  0   1  1  1  1 |   0  3   1  1  1   1  1  4  0
End of individual info.


No ancenstral population recorded.
>>>
```

4.3 Writing pure Python operator

Now we know how to access information for individuals in a population, but how can we use them in reality? Namely, how can you write an pure Python operator?

4.3.1 Use `pyOperator`

There are two kinds of pure Python operators. The first one is easy: define a function and wrap it with a `pyOperator` operator. This method is highly recommended because of its simplicity. Many user scripts will use this kind of pure Python operator. You can find such examples in scripts directory. A good one may be simuCDCV.py where a pure Python operator is used to calculate and visualize special statistics.

For example, if you would like to record a silly statistics, namely the genotype of the m individual at locu n, you can do:

```
def sillyStat(pop, para):
  # para can be used to pass any number of parameters
  (filename, m, n) = para # unpack parameter
  f = open(filename)
  f.write('%d ' % pop.individual(m).allele(n) )
  f.close()
# then in the evole function
evolve(...
  ops=[ # other operators
    pyOperator(func=sillyStat, param=('file.txt', 2, 1) )
  ]
)
```

`pyOperator` is by default a post-mating operator, you can redefine its stage by `stage` parameter.

4.3.2 Use Python `eval` function

This kind of pure Python operators acts more like an ordinary operator. They are usually `pyEval` or `pyExec` operators returned by a wrapper function. For example, the following function defines a `tab` operator:

Example 4.4: Tab operator

```
>>> def tab(**kwargs):
```

```
...     parm = ''
...     for (k,v) in kwargs.items():
...         parm += ' , ' + str(k) + '=' + str(v)
...     cmd = r'output( """\t""" ' + parm + ')'
...     # print cmd
...     return eval(cmd)
... #end
...
```

This function actually returns an operator

```
output(r"\t", rep=REP_LAST, begin=500)
```

This kind of operators have some advantages, namely

- it acts more like ordinary operator.
- it is more efficient since it is handled (at least the first layer) by a C/C++ operator.

However, because of its complexity, such operators can only be found in system modules. You can ignore the rest of this section if `pyOperator` is enough to you.

To define a pure Python operator, here are what you will generally do:

- write a function that acts on a population. This function should be able to be called like `func(simu.population(0))`.
- wrap this function as an operator.

For example, function `saveInFstatFormat(pop, output, outputExpr, dict)` saves a population in FSTAT format. Its definition is (first 15 lines)

Example 4.5: genotype

```
>>> print ind.sex()
2
>>> print ind.sexChar()
F
>>> ind.setSex(Female)
>>> ind.setAffected(True)
>>> print ind.tag()
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
AttributeError: 'individual' object has no attribute 'tag'
>>> ind.setTag([1,2])
Traceback (most recent call last):
  File "refManual.py", line 1, in ?
    #
AttributeError: 'individual' object has no attribute 'setTag'
>>> Dump(pop)
Ploidy:                 2
Number of chrom:        3
Number of loci:         2 3 4
Maximum allele state:   65535
Loci positions:
```

```
                1 2
                0.5 2 3
                1 2 3 4
  Loci names:
                loc1-1 loc1-2
                loc2-1 loc2-2 loc2-3
                loc3-1 loc3-2 loc3-3 loc3-4
  population size:        1
  Number of subPop:       1
  Subpop sizes:           1
  Number of ancestral populations:        0
  individual info:
  sub population 0:
     0: FA   1  0   1  1  0   1  1  1  1 |   0  3   1  1  1   1  1  4  0
  End of individual info.


  No ancenstral population recorded.
  >>>
```

Note that

- you can use this function independently like

```
saveInFstatFormat(simu.population(1),'a.txt')
```

- `pop.vars()` is used to evaluate `outputExpr`.

Then you can wrap this function by an operator, actually a function that returns a `pyEval` operator:

Example 4.6: save fstat

```
>>> def saveFstat(output='', outputExpr='', **kwargs):
...     # deal with additional arguments
...     parm = ''
...     for (k,v) in kwargs.items():
...         parm += str(k) + '=' + str(v) + ', '
...     # pyEval( exposePop=1, param?, stmts="""
...     # saveInFSTATFormat( pop, rep=rep?, output=output?, outputExpr=outputExpr?)
...     # """)
...     opt = '''pyEval(exposePop=1, %s
...         stmts=r\'\'\'saveInFstatFormat(pop, rep=rep, output=r"""%s""",
...         outputExpr=r"""%s""" )\'\'\')''' % ( parm, output, outputExpr)
...     # print opt
...     return eval(opt)
... #end
...
>>>
```

This function takes all parameters of an ordinary operator:

```
saveFstat(at=[-1], outputExpr=r'a'+str(rep)+'.txt')
```

and generates a `pyEval` operator (use above example).

```
pyEval(exposePop=1, at=[-1],
  stmts=r"""saveInFSTATFormat(pop,
    output='''''', outputExpr=r''' 'a'+str(rep)+'.txt' """
  )
```

In this example,

- `pyEval` works in the local namespace of each replicate. To access that replicate of population, you should use the magic parameter `exposePop` of `pyEval`. When set `True`, `pyEval` will automatically set a variable `pop` in the current local namespace before any statement is executed. This is why we can call `saveInFSTATFormat(pop...)`
- `"'a"'` quotes are used to avoid conflicts with quotes in `outputExpr` etc.

.

4.4 Ultimate extension: working in C++

It is sometimes desired to write simuPOP extension in C++. For example,

- when you need some other mating scheme.
- when you need certain operator that a pure Python implementation would be too slow.
- if some aspect of simuPOP is too limited (like the number of maximum alleles).

It is not difficult to write simuPOP extension in C++, once you know how simuPOP is organized. The general procedure is

- install the latest version of SWIG (>1.3.28)
- check out simuPOP source using subversion
- build from source and see if your programming environment works well
- to add an operator, make changes in appropriate `.h` file. Check `simuPOP_common.i` if your operator can not be used.

The source code is reasonably well commented with full doxygen based documentation. Please post to the simuPOP forum if you encounter any problem while writing operators in C++.

4.5 Debugging

4.5.1 Test scripts

There are many test scripts under the `test` directory. It is recommended that you run the test scripts after you installed simuPOP. This will make sure that your system is working correctly. To run all tests, run

```
sh run_tests.sh
```

Or, if you do not install RPy and R, run

```
sh run_tests.sh norpy
```

Please report any failed test.

4.5.2 Memory leak detection

Python extensions tend to have memory leak problem, caused by the refcount mechanism. If your simuPOP script uses more and more RAM without population size increase, you may have this problem. You may try to disable individual operators and find out the offending operator if the problem persist.

Potential simuPOP developers can make use of simuPOP's built-in refcount detection mechanism. To use it,

- compile Python with configure option – with - `pydebug` . This will enable `sys.totalrefcount()` etc.

- compile simuPOP with - `DPy_REF_DEBUG`. This can be done in setup.py, or better in SConstruct.

`simulator.evolve` will check reference counts at the end of each generation and report any increased reference count. Some operators may create Python objects (like ascertainment operators) but if you see repeated warnings at each generation, there is definitely a memory leak.