Introduction

The AgeStructureNb GUI interface offers a user interface to allow easy access to simuPOP-based simulations[1] and the LDNe[2]-based Nb and Ne estimations as implemented in Tiago Antao's python program, AgeStrucureNe, available at https://github.com/tiagoantao/AgeStructureNe.git. We also offer an interface for plotting Nb and Ne estimations, and regressions based on the estimations. The original analyses based on Tiago's code, with their applications to many species, are available in several publications, including [refs to AgeStructureNe - based pubs].

Our program offers a separate interface to perform three functions: population simulation, Nb and Ne estimation, and estimate visualization. The genepop file output from a simulation can be loaded into an Nb estimation interface, and in turn, the output from an Nb estimation can be loaded into a visualization interface. The Nb estimation interface can also use any genepop file for input.

---

[1] Bo Peng and Marek Kimmel, "SimuPOP: A Forward-Time Population Genetics Simulation Environment," *Bioinformatics* 21, no. 18 (September 15, 2005): 3686–87, doi:10.1093/bioinformatics/bti584.

[2] Robin S. Waples and Chi Do, "Ldne: A Program for Estimating Effective Population Size from Data on Linkage Disequilibrium," *Molecular Ecology Resources* 8, no. 4 (July 1, 2008): 753–56, doi:10.1111/j.1755-0998.2007.02061.x. Note that our program uses the beta version 2, the source code generously supplied to us by the authors.

Installation

The program can be downloaded from https://www.github.com/popgengui/negui. Click on the button labeled "clone or download". You can keep the program in any directory to which it can be written. For dependency details and other installation advice, see the README.md file provided with the program files.

Starting the program

The program is launched using a python 2.7 or python 3.6 executable, invoking the negui.py module. Please see the README.md file for details on different ways to start the program.

Loading interfaces

To load one or more of the three interfaces for performing simulations, Nb/Ne estimations, or plotting programs, from the main menu click on the "New" menu (fig. 1). You can load any number of interfaces and run them simultaneously, though running too many at once can tax your computers cpu and/or memory capacity to a standstill.
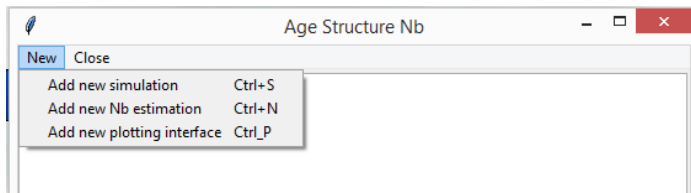
Figure 1 Adding an interface

Running a simulation
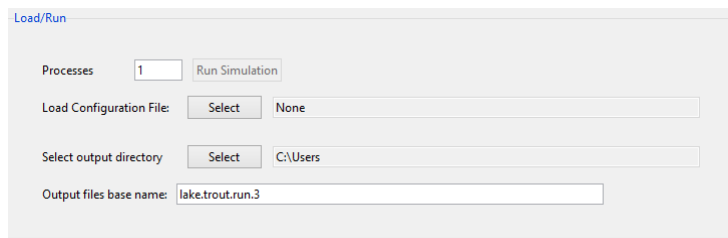
a. Load a simulation interface with the add menu (Figure 1), and set the parameters with the provided controls. Steps for preparing the interface to run a simulation follow.

b. Load a configuration file. The initial simulation interface requires the user to load a configuration file (Figure 2). The user can load a configuration file provided with the program, found inside the configuration/simulation directory inside the main program directory. Note that you can open one of these files and change the parameters manually, if you prefer it to setting them in the interface

c. Adjust the simulation parameters. With a configuration file loaded (Figure 3) you can change the values in the editable controls. These are detailed below in the "Simulation Input" section.

d. Click the button labeled "run simulation," and the simulation will start. The button's text now changes to say "cancel simulation," and next to it a new label notes that a simulation is in progress. While the simulation is in progress, the parameter controls are disabled.

Simulation Input

The simulation inter face has 5 subframes that divide the parameters by category.
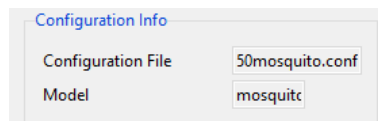
a. **The Load/Run section** (Figure 2) of the Simulation interface offers parameters related to input and output files.

Figure 2 Simulation interface, Load/Run section.

   i. **Processes**, with valid values between 1 and the total number of virtual cores in your computer. Multiple processes are only useful if you have set the "Replicates" parameter (see the Simulation subframe details below) to a value greather than one.
   ii. **Configuration File**. Press the Select button next to the label, *Load Configuration file* to load a configuration file into the interface. We have included configuration files for many species. These can be found in the "configuration_files/simulation" path inside the main program folder.
   iii. **Output directory**. Press the select button next to the label, *Select output directory*, select a folder for the output files written by the simulation
   iv. **Output files base name**. You can type in a base name for the simulation output files. The simulation will prepend this to the *.genepop, *.conf, *_age_totals.tsv and *_nb_values.tsv output files (see the "Simulation Output" section below).

b. **Configuration Info section** (Figure 3). This group simply shows you the input file information and has no settable parameters.

Figure 3 Simulation interface, Configuration info

i. **Configuration File Name**. This gives the file name of the loaded configuration file.
ii. **Model Name**. This gives the name of the model parameterized by the configuration file. In our example configuration files, the model name is usually a species' common name.

Figure 4, Simulation interface, population section

c. **The Population section** (Figure 4) offers many parameter settings that characterize the population's size and fecundity.

i. **N0 (Newborns)**. This gives the number of newborns added at each simulated reproductive cycle. This value is not editable directly, but is calculated using several values, all of which are editable. These including **Nb**, **Nb/Nc** , **Female, Male Survival** , and the probability of male birth. The N0 is recalculated whenever any of these values changes, using the following procedure:
    1. Assign an Nc value, as **Nb** divided by **Nb/Nc** .
    2. Assign a current_male_proportion equal to the **Probability of male birth** .
    3. Assign a current_female_proportion equal to 1 – the **Probability of male birth** .
    4. Assign a cumulative_proportion=1.

5. For each age value age_val giving a male and female survival rate:
   a. Update, current_male_proportion=current_male_proportion x male_survival at age_val.
   b. Update, current_female=current_remale x female_survival at age_val.
   c. Update cumulative_proportion =cumulative_proportion + current_male_proportion.
   d. Update cumulative_proportion=cumulative_proportion + current_female_proportion.
6. Set **N0**=Nc/cumulative_proportion, rounding it to the nearest integer.

ii. **Nb/Nc** is the effective number of breeders in one reproductive cycle divided by the cohort size.

iii. **Nb/Ne** is the ratio of the effective number of breeders in one reproductive cycle to the effective population size per generation. This value is not used in the simulation itself, but is written to the output genepop file, and can be used in the Nb estimation interface to make a bias correction in the LDNe estimation of Nb (see the section "Nb estimation input," below).

iv. **Nb** is the effective number of breeders in one reproductive cycle.

v. **Nb Tolerance** is the proportion of the **Nb** parameter by which a new population is added to the simulation after at each reproductive cycle. For example, if the **Nb** is set at 600, and the **Nb Tolerance** is set at 0.02 then populations created at each reproductive cycle (after the burn-in period, explained below), must have an **Nb** value, as calculated using the parentage analysis without parents (PWoP) procedure[3].

vi. **Ages** gives the number of age classes for the population to be simulated. Note that this is disabled, and that the length of the lists for **Female, Male Fecundity** and **Female, Male Survival** values (see below) are set to length **Ages** minus one for the former and **Ages** minus two for the latter. The age value and changes in these lists, therefore, need to be edited in either a life table or configuration file (see the section "Manually editing life tables and configuration files."

vii. **Female, Male Survival** are lists whose i^{th} item gives the probability of survival for an individual of the ith age category.

viii. **Female, Male Fecundity** are lists whose i^{th} item gives the probability of reproducing for individuals of the ith age category.

ix. **Force Skip** gives a probability, for each non-zero value, f_a, in the female fecundity list, that during a given reproductive cycle r, the value will be replaced with zero. Such replacement means that females belonging to the age class a, given by f_a, for cycle r, are infertile. This parameter is set (assigned a non-zero value) in only a few of the configuration files we copied from the AgeStructureNe program, and we have not made it editable in our interface. In your own custom configuration files you can set it to any value 0 through 100 (the value shown in the interface will be the file's value divided by 100).

x. **Litter**, if not a "None" value, will be a list of integers, affecting litter sizes. Note that we do not allow interface editing of these parameters, but note that, as above for the **Force Skip** setting, you can inter this parameter value in a

[3] Robin S. Waples and Ryan K. Waples, "Inbreeding Effective Population Size and Parentage Analysis without Parents," *Molecular Ecology Resources* 11 (March 1, 2011): 162–71, doi:10.1111/j.1755-0998.2010.02942.x.

Comment [TC1]: This definition needs checking.

Comment [TC2]: More detail about how the sim uses this value

Comment [TC3]: Also need more detail about role in sim.

configuration file. This should be a list, and can have one of 2 valid configurations:

    a. The list can have a single value l, and $l < 0$, then at each reproductive cycle the maximum possible number of offspring available to each reproducing female is given by $l * -1$.

    b. Otherwise, the list should have (positive) integers. In this case these integers proportionally allot litter sizes, as given by their indices in the list. In particular, at each reproductive cycle, as a female is chosen to mate:

        i. An age a is chosen randomly.

        ii. A female f_a is chosen randomly from the females of age a.

        iii. A list index i (i.e. one of $1,2,3…n$, where n is the number of items in the *litter* list), is selected by weighted probability, proportionally according to the ratio of each list value to the sum of the list values.

        iv. Female f_a is then the mother of the next i offspring (i.e. the female selection steps are skipped for the next i parings, since f_a is the female of the pair.).   Thus, she will parent the next i offspring, unless the *jth* of her offspring assignments produces the maximum total offspring for the cycle (i.e. **N0** is reached), and $j < i$.

xi. **Reproductive cycles** shows the total number of reproductive cycles that will be simulated.

xii. **Monogamous**, when checked, tells the simulation to enforce monogamy.

xiii. **Probability of male birth** is used during reproductive cycles to determine the sex of new individuals. As noted above in the description of the *N0*, it is used to in the **N0** calculation, and so the latter is recalculated when this value is changed. When the **Cull method** is set to *equal_sex_ratio*, this parameter is automatically set to 0.5, and it's entry box is disabled.

xiv. **Population size** shows the number of individuals that will be created in the simulation's initial population. Thereafter the size will change according to the reproductive parameters, notably *N0*.

d. **The Genome section** (Figure 5) parameters determine the simulated individuals' allelic content.

Figure 5 Simulation interface, genome section

i. **Mutation frequency**, if non-zero, is applied to microsatellites (not to SNPs). It will be used to set the simuPOP simulation StepwiseMutater's rate parameter.

ii. **Number of microsatellites**, simulated as diploid. Note that in cases in which you specify both *n* microsatellites and *m* SNPs, in the output genepop file, the first *n* loci are the microsatellites and the last *m* loci are the SNPs.

iii. **Number of SNPs**, simulated as diploid.

iv. **Starting Msat allele total** gives the initial number of microsatellite alleles present in each microsatellite in the initial population. For each microsatellite, the initial genotype frequencies are drawn from the Dirichlet distribution. With **Number of microsatellites** set to 10, for example, each microsatellite will have 10 alleles with frequencies given by the Dirichlet distribution of order 10, with alpha's uniformly set to 1.0. The maximum allowed number of alleles is 100.

Figure 6, Simulation interface, simulation section

e. **The Simulation section** (Figure 6) parameters determine several per-cycle behaviours:

i. **Cull method** indicates one of two possible per-cycle methods whereby individuals are removed from the population.

> **Comment [TC5]:** This is best explained after a consult with Brian T

ii. **Replicates** sets the number of independent simulations run with the current parameter set. These can be run in parallel if you specify more than one process in the **Processes** parameter.

iii. **Nb and census adjustment** offers one or more specifications that will change the target Nb and the number of individuals in the population by a fixed rate and at a range of cycles (one or more). Entries are of the form min-max:rate, specifying a change in Nb and census size applied at cycle numbers min through max. The values conform to min <= 2 <= max, and rate >= 0.0. No adjustment is made with rate = 0.0. For example, to reduce the Nb and the total number of individuals by a tenth at cycle 3, you would edit the entry to read, 3-3:0.1. The adjustments are different, depending whether the rate is less than or greater than 1.0.

1. If the rate is less than 1.0, the target *Nb* value, and each age class in the current census is reduced by the proportion given by rate. Note that the change in *Nb* will result in a change to *N0* as described above the Population section's description of *N0*.

2. If the rate is more than 1.0, the target Nb value will be multiplied by the rate, with a resulting recalculation of *N0*. No change will be made to the current census.

iv. **Skip breeding probability**, if its value is not "None," should be a list of percentages. It effects the number of available females of a given age at a given cycle number *c*. The *ith* percent p_i gives the probability (p/100) that a female of age=*i*, is not able to breed in cycle *c*. Like the **Litter** and **Force Skip** parameters, this parameter is not settable in the interface, but can be included in your configuration file.

v. **Cycles of burn-in**, give an integer *n* in the range $1 <= n <= r$, with *r* giving the total **Reproductive cycles** . This value tells simulation that the Nb tolerance test (see the **Nb Tolerance** parameter description) should not be performed for the first *n* cycles. The default value for this parameter equals the number of **Ages** in the model.

vi. **Start recording at cycle number** *c* will result in the genepop file containing only the populations of cycles c through *r*. *r* = total **Reproductive cycles** . This can greatly reduce the size of the output genepop file, when you are interested only in the last *r-c* cycles, but want to simulate many cycles before recording, and/or have large populations/loci to simulate.

Simulation output

When a simulation is complete the message "simulation in progress" will disappear from the intervace and the its editable entry boxes will no longer be grayed-out. A completed simulation delivers a genepop file for each replicate, named using the *output base name* parameter shown in the *Load/Run* section of the Simulation Input, the base name extended with a replicate number *n* and a "genepop" extension, so that, for example, if your simulation output base name is bulltrout, and you specified 3 replicates, the output file for the 3rd replicate would be named "bulltrout.r3.genepop." Also, there are three files produced during the first replicate only, all prefixed with the *output base name*. One with the extension "conf," lists the parameter settings for the simulation (and, hence, for all replicates), another has extension "_age_counts_by_gen.tsv," and a third file has extension "_nb_values_calc_by_gen.tsv". Details on the output files follow.

1. **The conf file** shows the parameter settings used in the simulation (except the number of replicates, which it always sets to one). This file can be loaded into another instance of the Simulation Input (see the *Load/Run* parameter) and another simulation with matching parameters can be run. Conveniently, if it represents many customized settings on a former configuration file, small changes to it can be made to run a simulation similar, but without having to re-enter all of the settings used to create it.

2. **The age counts file** is a table with tab-delimited fields that gives a count of total individuals for each age class, for each reproductive cycle. The first line in the file gives column headers, the first "generation," referring to reproductive cycle number, a zero-based count of reproductive cycles, and the rest listing age classes simply as 1,2,3…*t*, *t* = total age classes. This file is created only for the first simulation replicate.

3. **The Nb values file** is a table with tab-delimited fields giving the PWoP[3] Nb values calculated during the simulation, and used to compare to the target **Nb** value +/- the **Nb Tolerance** value. The first column gives the zero based reproductive cycle number and the second the PWoP-based Nb value that passed the tolerance test, and represents the

accepted population for that cycle. This file is created only for the first simulation replicate.

4. **The genepop file** conforms to the genepop file standards given at http://genepop.curtin.edu.au/help_input.html. The header line notes the name of the *.gen file it came from, which simply names an intermediate file from which it derived it's population information. It also gives the value of **Nb/Ne**. It the value is non-zero, it can be loaded automatically into the *Nb/Ne estimation interface* (see the *Parameters section* of the *Nb/Ne Estimation* interface description. The second line of the genepop file gives the name of the first loci, which is simply "l0." Each consecutive loci, l0, l1,l2…L-1 (where L gives the total number of microsatellites plus the total number of SNPs) is listed on a separate line. Note that the first M loci will represent the microsatellites, and the last S loci will represent the SNPs, with M and S the totals given in the *Genome* section of the Simulation Input. Thereafter the file consists of separate "pop" sections, each representing a reproductive cycle. The first n cycles (as numbered 1,2,3…n) will not be in the file if the *Start at cycle number* parameter is set to $n + 1$. The population for each cycle is listed, in order of cycle number. Each is demarked by a line with "pop" as its sole entry. Individuals, one to a line, follow each "pop" entry. Each individual as an ID with multiple fields delimited by a semicolon, giving, <individual id number>;<sex (1 = male, 2 = female)>;<id of father>;<id of mother>;<age class>. These are followed by a comma, and then a space-delimited set of alleles for each locus named in the lines 2 – total number of loci. Note that these allele entries represent diploidy, and use 3-digit allele numbering so that, for each loci, allele one is named by the first 3 digits, and allele 2 by the last 3.

Running an Nb or Ne Estimation

Figure 7 Nb/Ne estimations interface, Load/Run section

The Nb (and Ne) estimation interface performs and LDNe[2] based Nb or Ne estimation from genepop file input as supplied by the user. While it was developed in concert with the simuPOP-based simulation output from our program's interface, it will perform estimations based on any genepop file conforming to the format used by our simulation-based files. To run estimations:
.

a. Load a "new nb interface" with the add menu (Figure 7), and set the parameters with the provided controls.

b. Load a genepop file as described in the *Simulation output* section.

c. Adjust the estimation parameters.   The parameters are detailed in the *Nb/Ne estimations input* section below.

d. Click the button labeled "Run Nb Estimation," and the computations will start. The button's text now changes to say "cancel simulation," and next to it a new label will note that "estimations in progress. As in the other interfaces, while the estimations are in progress, the parameter controls are disabled.

Nb/Ne Estimations input

The interface provides for multiple subsampling schemes of both individuals and loci within the input genepop file pop sections. The sub-sections of the interface follow.

a. **The Load/Run section**   (Figure 7) offers an interface to load input and name the output files.
   i.   **Total processes**. The program will run estimations on the individual "pop" sections in parallel using the number of processes set here. It is usually advisable, unless your computer has man process already running, to use most if not all of your available (virtual) processing cores, to speed up the estimation runparameter.
   ii. **The Load genepop files button** offers a file loading interface to loacate and load one or more **genepop file**(s). Note that when you load multiple genepop files, the parameter settings will be applied to all. In particular, activating an **Nb bias adjustment** will apply it to all the files, so that only data with which it is compatible should be loaded. This also applies to other parameters, such as population number range [bookmark] and loci number range [bookmark].
   iii. **Select output directory**.   By clicking on the button and choosing your preferred folder you select where the estimation output files will be written. Note that this will also be used as a temporary directory in which intermediate files will be written inside new directories with the "tmp" prefix, ending in random characters. These files will be removed on completion of the simulation. Sometimes, if the estimation run is cancelled or otherwise is interrupted, they will not be removed, but can be manually deleted from your directory.
   iv. **Output files base name**. The text entered here will become the prefix for the output files [bookmark].
b. **The Genepop Files Loaded section** (Figure 8) has a single box that shows you the names of the loaded genepop files. It is not an editable section.
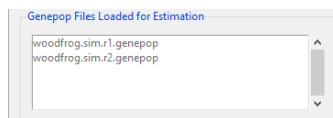
Figure 8, Nb estimation interface, genepop files loaded section

c. **The Parameters section** gives the main parameters, including the choice of subsampling in pop sections and/or loci.
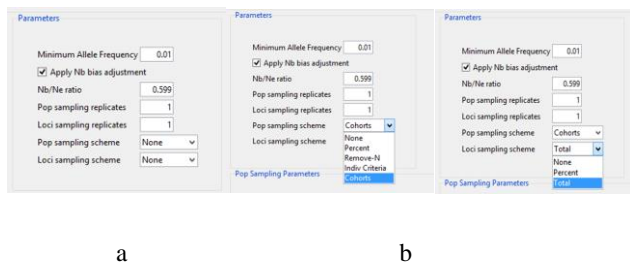

a   b   c

Figure 9, Nb estimations interface, parameters section. Subfigure a shows the section with no subsampling selected. Subfigure b shows the types of subsampling available for pop sections, and subfigure c shows those available for the loci.

i. **Minimum allele frequency** sets the threshold below which the LDNe program will ignore an allele in its LDNe calculation.

ii. **Nb bias adjustment** check box allows you to apply a bias adjustment to the estimations, as described in. Note that when you load a **genepop file** generated by the simulation interface, when check the bias adjustment box, the program will load the **Nb/Ne** value as set in the simulation interface. You can accept it or enter another value. If no value is available in the genepop file, then you will need to enter a non-zero value to make any bias adjustment.

iii. **Nb/Ne ratio** is that which will be used for the bias adjustment, when it is checked. A zero value or an un-checked box means no bias adjustment will be done.

iv. **Pop sampling replicates.** When set to *n*, for each estimate, for each pop sampling parameter, the estimate is repeated *n* times. While you can set this to any value for any subsampling scheme, note that it is sensible only when your subsampling parameter involves a random sample of individuals (see [bookmarks] schemes), or you have more than 1 loci subsampling parameter (see [bookmarks] loci subsampling schemes, or both. If there is no random subsampling, the replicates will be performed, with identical results.

v. **Loci sampling replicates**. When set to *n*, for each pop sampling replicate, for each loci subsampling parameter value, the estimate is repeated *n* times. As with the **Pop sampling replicates** parameter, if the loci subsampling scheme has no randomized subsample (see [bookmarks]), then the estimates will be identical.

vi. **Pop sampling scheme**. This drop-down box offers the following subsampling schemes (Figure 9b).
   1. *None*. This scheme uses all individuals with the pop section for the Nb or Ne estimation, unless the [bookmark for the max pop size] reduces the pop size.

2. *Percent*. When you select this scheme the [bookmark for pop sampling parameters] shows a "percent" box with two buttons below it (Figure 10b), "Add Value," and "Trim." You can edit any box currently in the list. Clicking "Add Value" will append a box to the list, its default value taken from its nearest neighbor. For each percentage p in the list, each pop section will be reduced to p percent of its total individuals, unless its census is not in the range given by [bookmark for min pop size] and [bookmark for max pop size], in which case the population will be skipped, and a message written to the [bookmark to msgs output file]. For each subsample the individuals are randomly selected. Also, for each p, an estimate is calculated $p*q*r$ times, where p equals **Pop sampling replicates**, q equals the number of pop loci subsample parameters (e.g. [bookmark loci subsample percentages] and r equals **Loci sampling replicates**.
3. *Remove-N*. This scheme also offers an editable list, which behaves as described for the *Percent* scheme. For this scheme, when you enter an integer N, the pop subsample will be that given by its total minus N, the removed N individuals randomly selected (except when $N = 1$, described below). As described for the *Percent* scheme, for each value N in the list, the estimate will be replicated $p*q*r$ times, except in the case of $N = 1$, in which case each estimate will be replicated $t*q*r$, where t equals the total individuals in the pop section. This means that all remove-1 cases will be estimated. Pop sections are skipped if the total individuals in the population are not in the range [bookmark for min pop size] and [bookmark for max pop size].
4. *Cohorts*. Because this scheme selects individuals by age, it requires input genepop files produced by the Simulation Input. When "cohorts" is selected the [bookmark to pop sampling params section] shows, besides its usual pop number and min/max pop size parameters, an entry labelled "Indiv max age (Figure 10). For each population, individuals outside the ages given by [0.0, max age] are excluded. However, because the simulations produce genepop files with no individuals aged less than 1.0, the interval is effectively [1.0, max age]. For each pop, and for each percentage value in the list, subsampling steps are:
   a. For each age value in [0.0, max age], count the total individuals in the age group.
   b. Find the smallest of the age group totals, t, from (1).
   c. Randomly subsample t individuals from each age group to get total sample size s.
   d. If s is within [min pop size, max pop size] then, for each percentage p in the percentage list, randomly select p percent of the collected individuals. If s is outside the range given by [min pop size, max pop size] an error occurs and the analysis is terminated.
5. *Individual Criteria*. [description]
ii. **Loci sampling scheme**. [description]
   1. *None*. [description]
   2. *Percent*. [description]

3. *Total*. [description]
b. **Pop sampling parameters section**. In this section (Figure 10) you set the pop section sampling parameters, which are presented according to the scheme selected in the **Pop sampling scheme** parameter.

Figure 10, pop sampling parameters section, as presented for. (a) the **None** *scheme, (b) the* **Percent** *scheme, (c)* **Remove-N**, *(d)* **Cohorts**, *and (e)* **Individual Criteria**.

i. **Pop number start**. When this is set to integer *n,* the estimates will skip pop section numbers (as ordered in the genepop file) in the range [1, *n* − 1].
ii. **Pop number end**. If the genepop file has *t* total pop sections, then, when this parameter is set to integer *n*, the estimates will skip pop section numbers in the range [*n* + 1, *t*].
iii. **Indiv min per pop**. [description]
iv. **Indiv max per pop**. [description]
v. **Scheme-specific parameters**. [description]
c. **Loci sampling parameters section**. [description]

Figure 11, locii sampling parameters section, as presented for the schemes (a) **None**, *(b)* **Percent**, *(c)* **Total**.

      i.   **Loci number start**. [description]
    ii.   **Loci number end**. [description]
  iii.   **Min Loci count**. [description]
  iv.   **Max Loci count**. [description]
    v.   **Scheme specific parameters**. [description]

Nb estimation output

a.   **Messages file** (*.msgs) [description]
b.   **Estimates table file** (*.tsv) [description]

Visualization Input

a.   **Load/Run section**. [description]
b.   **Tsv file loaded section**. [description]
c.   **Viz type section**. [description]
d.   **Regresssion plotting section**. [description]
e.   **Subsample plotting section**. [description]