Tracking somatic drift reveals small effective population size of stem cells and high somatic mutation rate in asexual planaria

Hosseinali Asgharian[1*], Joseph Dunham[3], Paul Marjoram[2], Sergey V. Nuzhdin[3]

[1] Department of Biochemistry and Biophysics, School of Medicine, University of California at San Francisco, San Francisco, California, United States of America

[2] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America

[3] Program in Molecular and Computational Biology, Dornsife College of Letters, Arts, and Sciences, University of Southern California, Los Angeles, California, United States of America

Sergey Nuzhdin

Phone: (213) 740-5773

Email: snuzhdin@usc.edu

Abstract

Planarian flatworms have emerged as highly promising models of body regeneration due to the many stem cells scattered through their bodies. Currently, there is no consensus as to the number of stem cells active in each cycle of regeneration or the equality of their relative contributions. We approached this problem with a population genetic model of genetic drift. We modeled the fissiparous life cycle of asexual planarians as an asexual population of cells that goes through repeated events of splitting into two subpopulations followed by population growth to restore the original size. We sampled a pedigree of obligate asexual clones of *Girardia cf. tigrina* at multiple time points encompassing 14 generations. Effective population size of stem cells was inferred from the magnitude of temporal fluctuations in the frequency of somatic variants and under most of the examined scenarios was estimated to be in the range of a few hundreds. Average genomic nucleotide diversity was 0.00398. Assuming neutral evolution, the somatic mutation rate was estimated in the $10^{-5} - 10^{-7}$ range. The surprisingly small effective number of propagating stem cells might contribute to reducing reproductive conflicts in clonal organisms.

Introduction

Planarian flatworms are a fascinating model system for studying body regeneration. After injury they can reconstruct their body from very small pieces of tissue; and, can grow or "degrow" by regulating the number of cells in their bodies in response to nutrient availability[1,2]. The asexual species in the clade are fissiparous[3,4]: the grown worm splits down the middle creating a head piece and a tail piece; each half is restored to a complete body through positionally regulated cell multiplication, cell death and differentiation[5]. Growth and regeneration rely mainly on a very large number of stem cells (neoblasts) which comprise 25-30% of the cells in a planarian's body by morphological observation[2]. Neoblasts are the only dividing cells in Planaria[2] but BrdU labeling experiments suggest that only a fraction of them are active at any given time[1]. Due to the heterogeneity of morphological and cellular features in stem cells, as well as their partial similarity to early post-mitotic committed progeny cells, mitotic molecular markers such as PCNA and H3P or BrdU labeling cannot visualize stem cells completely and exclusively [1,6]. Recently, irradiation and transplantation experiments suggested the surface protein Tetraspanin as a reliable marker for isolation of pluripotent stem cells[7]. Nevertheless, methods involving BrdU injection, experimental wounding, or irradiation significantly alter the normal physiology of the animal. Consequently, a quantitative understanding of the number of active stem cells and their relative contributions at each cycle of tissue regeneration during natural growth and reproduction is still lacking.

Planarians are a curious case for evolutionary genetic studies, too. Because asexual planarians do not undergo the single-cell bottleneck of zygote, significant genetic heterogeneity exists within a single worm's body[8], which invokes competition among diverged cellular lineages[9]. Mutator alleles remain linked with the mutations they cause in these non-recombining genomes and somatic mutations are transmitted to future generations. Theoretically, deleterious mutations (if not completely recessive) could be eliminated at the cellular level locally before they reach a frequency that can affect organismal fitness, while beneficial mutations give the mutator lineage a competitive advantage. These dynamics predict a higher optimal somatic mutation rate in clonal organisms[10].

In this study we present a simplified population genetic model of the life cycle of an asexual planarian. In this model, somatic cells play the role of asexually reproducing individuals in an expanding population, which is the planarian body. Over time, this population doubles in size and splits into two subpopulations, i.e., the head and tail pieces. We tracked fluctuations in the frequency of somatic variants over >10 generations and applied the theory of genetic drift at the cellular level to estimate the effective size of the stem cell population ($N_{e,sc}$). We showed that the allele frequency spectrum of somatic variants in our model system is shaped more strongly by genetic drift (random fluctuations affected by population size) versus genetic draft (selection acting on tightly linked loci). Then, we proceeded to estimate somatic mutation rate from $N_{e,sc}$ and the observed nucleotide diversity (π) according to expectations of the neutral theory.

Results

We established a line of lab-reared asexual flatworms from a single individual. Cytochrome oxidase subunit 1 (COI) DNA barcode and proportion of total reads mapping to different Dugesiidae genomes both identified this lineage as *Girardia cf. tigrina*. Details are provided in Methods and Tables S1 and S2. In several years of maintaining this line, no instance of sexual reproduction was observed. For this study, the worm lineage was followed for 14 generations of splitting and regrowth. Genomic libraries were prepared from tail pieces after splits 2, 6, 8, 10, 12 and 14 and sequenced in 2-3 replicates (Fig. 1). The fastq files were trimmed, filtered and deduplicated using Trimmomatic and BBTool's Clumpify and evaluated by Fastqc. Details of read pre-processing are provided in Methods and Table S3. The core idea of this project is based on the measurement of temporal variance in allele frequencies caused by genetic drift. We needed to ascertain that the sample-to-sample differences were caused mainly by drift and not technical variability. One of the replicates from sample XIV failed, making samples II and XII the two samples farthest apart in time that had replicated sequenced libraries. The main population genetic inferences were made based on generation 2 and 12. Freebayes was run on individual replicates as separate samples first. Bi-allelic SNPs with coverage 10-40X in all samples were subjected to principal component analysis (PCA). PCA confirmed that the difference between replicates was indeed much smaller than that between biological samples (Fig. S1). FreeBayes was run a second time merging all the replicates of each biological sample. Allele frequencies of positions with coverage 10-60X from the merged-rep VCF were used for population genetic analysis. Mean sequencing coverage of SNPs was 12.1X for sample XIV and 19.9-36.6X for the other merged-rep samples.

Divergence from the reference genome. In the merged samples, divergence was calculated as the ratio of the number of positions in the coverage 10-60X range with $AAF > 0.99$ to the total number of positions covered 10-60X. Average divergence from the six samples was 1.112% (range 1.03-1.28%) which corroborates taxonomic identification of our specimens as conspecific or congeneric to *Girardia tigrina*.

Somatic drift vs. somatic draft. The allele frequency spectrum from samples II and XII is given in Supplementary figure S2. Assuming the reference allele to be ancestral in most positions, patterns of the alternative (derived) allele frequency roughly match the neutral expectation except for a peak of $f_{alt} \cong 1$ representing positions of fixed divergence from the reference. Simulations have shown that in the absence of recombination, strong selection on linked loci can distort the distribution of allele frequencies in a specific way: the density of derived allele frequency falls off much more steeply under linked selection (draft) than under neutral evolution (drift). Denoting derived allele frequency as ν and the corresponding density as $f(\nu)$, derived allele frequency falls like $f(\nu) \sim \nu^{-1}$ under drift but like $f(\nu) \sim \nu^{-2}$ under draft[11]. In our dataset, the $f(\nu) \sim \nu^{-1}$ model ($R^2 = 0.938$) fits better than the $f(\nu) \sim \nu^{-2}$ model ($R^2 = 0.832$) (See Table S4 for details). This suggests that: 1) The observed AFS (Fig. S2) is more consistent with drift than draft, although draft cannot be rejected and probably plays some role (0.832 is not much smaller than 0.938). 2) The drift model fits quite well, which validates the idea of estimating stem cell N_e based on somatic drift. Fig. S3 illustrates temporal fluctuation in alternative allele frequency of 9 randomly selected loci.

Effective size of stem cell population and somatic mutation rate: We calculated $N_{e,sc}$ according to Equations 1-3 (see Methods) between generations II and XIV ($i = 2, j = 14$) which encompassed the longest interval of generations, also between generations II and XII which covered the longest interval of generations where replicated sequencing data was available (Table 1). The estimate of $N_{e,sc}$ changes almost two-fold depending on the choice of final sampling generation (XII or XIV) which is expected due to the stochastic nature of somatic drift. It also varies almost linearly with the number of generations elapsed (t), which in turn depends on the ratio of cell to organismal generations (g). Although the rate of tissue turnover in planarians under natural physiological conditions has not been measured experimentally, a lower bound for g can be obtained by assuming that 1) a stem cell's proliferation rate is independent of its age, and 2) when the worm grows fast under favorable conditions, most cells are young, and therefore the rate of apoptosis is negligible compared to the rate of cell division. Under these assumptions, the number of cellular generations will be approximately half the number of organismal generations; because, after each split and regrowth event, half the cells come directly from the previous generation (number of generations elapsed = 0) and the other half are newly produced by stem cells (number of generations elapsed = 1), bringing the average over all cells to 0.5. Homeostatic tissue turnover becomes increasingly prominent the slower the worm grows because more and more somatic cells age, die, and are replaced, adding to the number of elapsed cellular generations. The smallest estimate is $N_{e,sc} = 71.1$ which corresponds to $g = 0.5$ based on allele frequencies from generations II and XII (Table 1).

Nucleotide diversity (π) for each sample was estimated as the product of nucleotide diversity at bi-allelic SNPs (π^*) and the proportion of polymorphic positions in the sample. Average nucleotide diversity across all samples was $\bar{\pi} = 0.00398$. Assuming mutation-drift balance under neutrality, the somatic mutation rate was calculated according to the haploid form of the equation $\pi = 2N_e\mu$ or $\mu_{som} = \frac{\overline{\pi}}{2N_{e,sc}}$. Estimates of $N_{e,sc}$ and μ_{som} under several scenarios are given in Table 1. The eight examined parameter sets estimate μ in the $2.7 \times 10^{-5} - 7.2 \times 10^{-7}$ range. These values are likely overestimates because nucleotide diversity shows a decreasing trend down the generations in our experiment (Fig. S4) indicating that the system may not be at drift-mutation balance and somatic genetic variation may not be in a steady state.

Discussion

In this study we modeled the cells in asexual planarians as individuals in an asexually reproducing population and used classical population genetic theory to estimate the effective number of stem cells and somatic mutation rate in a fissiparous strain of *Girardia cf. tigrina*. Visual inspection of somatic allele frequencies of nine example loci showed random fluctuations (Fig. S3). Evaluating the density of the allele frequency spectrum (Table S4) further corroborated the role of somatic drift at the cellular level. In the original scheme theorized by Waples, a population is sampled at two time points and an estimate of effective population size (N_e) is derived from observed temporal changes in allele frequencies[12]. To make the model more analytically tractable and free of certain restricting assumptions, they recommend a "sampling before reproduction" plan to ensure that there will be no overlap (and therefore no covariance) between the reproducing individuals (contributing to N_e) and individuals sampled for allele frequency estimation. Such a separation is guaranteed in our model system as tail pieces are sequenced while head pieces grow to create the next generation. The theory we used assumes discrete generations, but we know planarian cells do not divide and die synchronously. Results from applying the classical N_{e,F_k} method to populations with overlapping generations may be biased[13]. Accuracy of N_e estimation can be improved by examining more loci, using loci with intermediate allele frequencies, sampling more individuals and increasing the time between samples (at least 3-5 generations apart, preferably more)[14,15]. Following these guidelines, we estimated $N_{e,sc}$ from the longest interval between our samples that contains sequencing replicates. Our estimate of $N_{e,sc}$ comes from 370 loci which is much larger than recommended to achieve acceptable accuracy and precision in the references. We also omitted SNPs with minor allele frequency <0.1. One stipulation is that these loci are only semi-independent: due to lack of recombination, they cannot be considered independent from a genetic perspective; however, a mitigating factor is that by the nature of pooled sequencing of a genetically heterogenous tissue (planarian tail piece), reads covering different genomic positions come from chromosomes that likely originated from different cells. The situation is therefore as if different individuals from the population were sampled to measure allele frequencies at each locus.

Our estimates of the effective number of stem cells even under the most permissive scenarios are still much smaller than the number of stem cells suggested by microscopic observations (tens of thousands). It has been suggested that 20-30% of cells in a worm's body are stem cells[16–18]. A 1-cm long Dugesia was estimated to contain approximately 2×10^5 cells[19]. This gives an estimate of 40000-60000 stem cells which is 1 or 2 orders of magnitude higher than our estimate of a few hundreds to a few thousands

(Table 1). The most likely explanation is that a small fraction of stem cells, e.g., stem cells closer to the fission site, contribute disproportionately highly to regeneration. It should be noted that we estimated the number of active stem cells in the head piece, which is known to have fewer stem cells than the tail piece, with the area anterior to the eyes practically devoid of any stem cells. Underestimation of homeostatic tissue turnover rate reflected in g (ratio of cellular to organismal generations) is a possibility. Hopefully, experimental data in the future will quantify tissue turnover and g can be specified with more certainty. It has been shown that selection will not affect $\hat{F}_k$ much under plausible circumstances, especially when $\frac{t}{N_e}$ is small[20]. It has been suggested that variable selection and changes in demographical parameters can lead to overestimation of F_k and underestimation of N_e[12]; but these are unlikely to have played any significant role in our model system of worms reared in controlled lab conditions. If the effective population size (of stem cells or otherwise) varies from generation to generation, N_{e,F_k} estimates the harmonic mean of N_es across generations.

We have modeled the planarian body as a freely mixing pool of cells, in other words we have not incorporated body structure in our model. The analytical model used here to estimate N_e assumes random selection of *effective members* of the population to produce the next generation. Body structure may contribute to $N_{e,sc}$ being smaller than the microscopically observed number of stem cells, since stem cells closer to the fission site probably play a more important role in regeneration. However, it is unlikely to violate the assumptions of our model in the long term for two reasons: 1) After experimental amputation, natural fission, or during the processes of growth and degrowth, the worm's body undergoes extensive reshaping (known as morphallaxis) [2,21]. This means that body structure is likely not preserved from generation to generation, and therefore random activation of stem cells is an appropriate model for estimation of effective population size. 2) Irradiation-amputation experiments show that stem cells or their progeny can migrate long distances from their original position to the wound site to contribute to regeneration [1,2]. The reconstruction of the body after fission is not restricted to stem cells adjacent to the fission site although they might contribute more.

The estimated somatic mutation rate is of the order of $10^{-5} - 10^{-7}$ which is orders of magnitude higher than the norm in sexually reproducing eukaryotes (often in the $10^{-8} - 10^{-9}$ range). There are several points of consideration here. First, in most organisms, including humans, the rate of mutation in somatic tissues is about an order of magnitude higher than in the germline[22,23]. Second, theory predicts that mutation rate evolves to a minimum rate in sexual populations but could evolve to a non-minimal optimum in asexual populations under particular circumstances[24,25]. It has been shown that in the presence of strong selection among somatic cell lineages, a higher optimal mutation rate can evolve, because mutator alleles can benefit from the advantageous mutations they cause while deleterious mutations are eliminated before they get the chance to be transmitted to next generation[9,10]. The high level of somatic heterogeneity, also observed by Nishimura et al.[8], can provide the basis for such strong somatic selection. Third, the neutral expectation of $\pi = 2N_e\mu$ is derived assuming a steady state level of genetic variation (mutation-drift balance). However, nucleotide diversity shows a decreasing trend over time in our samples (Fig. S4) which permits a lower mutation rate than estimated from Equation 6. Most natural systems are expected to maintain a steady state level of genetic variation. We speculate that the observed decreasing trend in π is partly due to the fast growth and splitting of worms under favorable lab conditions. Under slower growth, more homeostatic tissue turnover would happen which would allow for the accumulation of as many new mutations as eliminated by somatic drift.

Methods

Worm collection and maintenance. The worms were collected from a stream in Almese, Italy on September 23, 2009. They were separated into individual vials and kept in standard rearing conditions and fed beef liver once a week followed by water exchanges. Over more than two years, the worms reproduced exclusively asexually though fissiparity. At the beginning of the experiment, a single lineage was followed for 14 generations of splitting and growth. After generations II, VI, VIII, X, XII and XIV, the tail piece was frozen for sequencing while the head piece was left to grow and further the lineage (Fig. 1).

Sequencing, QC and duplicate removal. Genomic libraries were produced from frozen tail pieces with 2-3 replicates according to the protocol described previously[26] and sequenced on Illumina HiSeq. Raw fastq files were examined by FastQC (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc*). Low quality segments were removed and short trimmed reads (<36 bases) dropped using Trimmomatic v.0.38[27] in the paired-end mode with the following options: */<path>/trimmomatic-0.38.jar PE <input files> <output files> ILLUMINACLIP:/<path>/adapters/TruSeq3-SE.fa:2:30:12:1:true LEADING:3 TRAILING:3 MAXINFO:40:0.4 MINLEN:36*. Only proper pairs (both mates surviving) were processed further. PCR duplicates were removed using the Clumpify command of BBTools (*https://jgi.doe.gov/data-and-tools/bbtools*) in paired-end mode with the following options: */<path> /clumpify.sh -Xmx230g <input files> <output files> dedupe=t reorder*. In addition to removing duplicates, Clumpify sorts fastq files for more efficient compression which reduces compressed file size and accelerates future processing steps. Trimmomatic and Clumpify output files were re-evaluated by FastQC.

Taxonomic identification. Based on initial morphological inspection, the specimens were tentatively assigned to family Dugesiidae. Molecular identification was carried out using cytochrome oxidase subunit 1 (COI) barcode sequences [28]. Pre-aligned COI barcodes of 72 species belonging to the order Tricladida including 16 species from family Dugesiidae were downloaded from the Barcode of Life database (boldsystems.org) public records. Whole genomic sequencing reads from two of our samples were mapped (separately) to the Tricladida COI sequences using bowtie2 (--end-to-end alignment default options). All our samples come from a single founder and belong to the same lineage; however, two samples were tested to ensure the reproducibility of identification. To verify the COI identification by a wider genomic scan, 1000 reads were randomly subsampled from each one of the 24 fastq files pertaining to the 12 paired-end sequenced samples and then pooled together. This pooled fastq was mapped to the concatenated fasta file comprising the four published Dugesiidae genomes on NCBI (*Schmidtea mediterranea* asexual strain CIW4, *S. mediterranea* sexual strain S2, *Dugesia japonica* and *Girardia tigrina*). Reads were mapped using bowtie2 end-to-end and local alignment options (both in the --very-sensitive mode). Although the samples had been sequenced paired-end, for taxonomic identification purposes alignments were performed separately for forward and reverse reads in the unpaired mode to avoid under-mapping of pairs due to insert size variation caused by indels and structural variations, which would not be uncommon in the taxonomic range of a large and diverse order such as Tricladida.

Mapping and Variant calling for population genetic analysis. Proper trimmed and deduplicated pairs were aligned to the *G. tigrina* genome using *bwa mem* with default options except setting the minimum seed length to 15 (*-k 15*) to facilitate the mapping of more divergent reads[29]. Variants were called using FreeBayes[30] with options *-F 0.01 -C 1 -m 20 -q 20 --pooled-continuous --use-reference-allele*. This configuration is recommended for variant calling from pooled-seq data with an unknown number of pooled samples (https://github.com/ekg/freebayes) (in our study, the exact number of cells per sample in unknown). The *--use-reference-allele* option ensures that the output includes positions where the base call in the samples differs from the reference allele even if they are monomorphic across the samples. The current *G. tigrina* assembly comprises >255k scaffolds. This created an error in the running of FreeBayes which we suspect was due to an internal algorithmic step concatenating the names of all chromosomes into a single string. To circumvent this problem, and since we needed only tens of SNPs for a reliable estimate of N_e, the bam files and the reference fasta files were filtered to contain only the longest contig (MQRK01218062.1, length=267531). The bam files were manually re-headered in two ways: 1. Rep-by-rep: A different sample name (bam header 'SM') was assigned to each replicate 2. Merged-reps: The same sample name (bam header 'SM') was assigned to all replicates of the same biological sample, the effect of which is to pool the reads from replicates during variant calling and the consequent calculation of allele frequencies. Sample XIV.8, for which only the reverse mate sequences were available, was excluded from variant calling and later analyses. Thus, all remaining biological samples except sample XIV were represented by at least two independently sequenced replicates. Coverage statistics of the re-headered bam files were obtained using the *samtools mpileup* function. VCF files were filtered using *vcftools* to keep bi-allelic SNPs only. A custom python code was run to keep only SNPs in the coverage range of 10-40X and 10-60X in rep-by-rep and merged-rep VCFs, respectively. Only samples from generations II and XII were used for population genetic analyses (see Results).

Consistency of allele frequency estimates among replicates was evaluated via principal component analysis.

Population genetic analyses. Divergence from the reference genome was calculated from each merged-rep sample as the ratio of SNPs covered 10-60X with alternative allele frequency $AAF > 0.99$ to the total number of positions covered 10-60X in the corresponding sample.

To compare the influence of drift and draft on the allele frequency spectrum, a vector of densities of allele frequencies for SNPs with AAF $0.01 - 0.9$ was obtained using the density() function in R. Denoting AAF as ν and its corresponding density as $f(\nu)$, we tested the goodness-of-fit of three linear models $f(\nu) \sim \nu$, $f(\nu) \sim \nu^{-1}$ and $f(\nu) \sim \nu^{-2}$ by their p-values and R^2 [11].

Effective population size of stem cells was calculated based on the change in allele frequencies of SNPs between generations using the theory laid out by Waples[12,31]. In this method, a parameter F_k is calculated from the transgenerational difference in allele frequencies the expected value of which depends on sample sizes at the initial and final sampling points (S_i and S_j), number of generations elapsed (t) and effective population size (N_e). We calculated F_k at each SNP as:

$$\hat{F}_k = \frac{\left(p_i - p_j\right)^2}{\frac{p_i + p_j}{2} - \left(\frac{p_i + p_j}{2}\right)^2} \qquad \text{Equation 1}$$

Where p_i and p_j are the frequency of the alternative allele at generations i and j, respectively (here: $i = 2, j = 12$). Only SNPs with AAF $0.1 - 0.9$ were included, to avoid biases introduced by very rare minor alleles[12]. The average F_k over loci ($\bar{F}_k$) was plugged into the following equation to obtain $N_{e,sc}$:

$$E(\bar{F}_k) = \frac{1}{S_i} + 1 - \left(1 - \frac{1}{N_{e,sc}}\right)^t \left(1 - \frac{1}{S_j}\right) \qquad \text{Equation 2}$$

Equation 2 can be rearranged to obtain $N_{e,sc}$:

$$N_{e,sc} = \frac{1}{1 - \sqrt[t]{\frac{1 + \frac{1}{S_i} - E(\bar{F}_k)}{1 - \frac{1}{S_j}}}} \qquad \text{Equation 3}$$

In the original formulation designed for a single locus, S_i and S_j are sample sizes at generations i and j [12]. Here, sample size is replaced by sequencing coverage. S_i and S_j are harmonic means of sequencing coverage across all examined SNPs in generations i and j, respectively. The number of generations elapsed between the two sampling points is represented by t. Since we are modeling cells as individuals,

t in our calculations must reflect cellular generations. As far as we know, the rate of tissue turn-over in planarians has not been measured quantitatively. We defined g as the unknown ratio of cellular generation / organismal generation and evaluated its effect on the calculated $N_{e,sc}$.

Nucleotide diversity (π) for each sample was calculated from positions covered 10-60X in that sample as:

$$\pi = \pi^* \times \frac{No.\ bi-allelic\ SNPs\ covered\ 10-60X}{No.\ positions\ covered\ 10-60X} \qquad \text{Equation 4}$$

Where π^* is nucleotide diversity at bi-allelic SNPs covered 10-60X and is calculated from the number of reference and alternative allele counts (Rc and Ac, respectively) as follows:

$$\pi^* = \left[\frac{Rc*Ac}{Combination\big((Rc+Ac),2\big)}\right]_{average\ over\ SNPs} \qquad \text{Equation 5}$$

The second term on the right side of equation 4 is the proportion of polymorphic positions. No restriction on initial AAF was imposed for calculation of π.

The somatic mutation rate was estimated according to the neutral theory expectations in haploid form:

$$\bar{\pi} = 2N_{e,sc}\mu_{som} \implies \mu_{som} = \bar{\pi}/2N_{e,sc} \qquad \text{Equation 6}$$

The haploid form of the neutral equation was chosen because in our Pool-seq data each genomic position in a sample is represented by individually sequenced chromosomes rather than diploid genotypes. Correspondingly, sample size at each position is set equal to the sequencing coverage at that position.

Acknowledgement

This work was supported by NIH Grant No. GM098741 and American Cancer Society Grant No. 130920-RSG-17-114-01-RMC.

References

1. Rink, J. C. Stem cell systems and regeneration in planaria. *Dev. Genes Evol.* **223**, 67–84 (2013).

2. Reddien, P. W. & Alvarado, A. S. Fundamentals of Planarian Regeneration. *Annu. Rev. Cell Dev. Biol.* **20**, 725–757 (2004).

3. Knakievicz, T., Vieira, S. M., Erdtmann, B. & Ferreira, H. B. Reproductive modes and life cycles of freshwater planarians (Platyhelminthes, Tricladida, Paludicula) from southern Brazil. *Invertebr. Biol.* **125**, 212–221 (2006).

4. Lázaro, E. M. *et al.* Molecular barcoding and phylogeography of sexual and asexual freshwater planarians of the genus Dugesia in the Western Mediterranean (Platyhelminthes, Tricladida, Dugesiidae). *Mol. Phylogenet. Evol.* **52**, 835–845 (2009).

5. Reddien, P. W. The Cellular and Molecular Basis for Planarian Regeneration. *Cell* **175**, 327–345 (2018).

6. Newmark, P. A. & Alvarado, A. S. Not Your Father'S Planarian: a Classic Model Enters the Era of Functional Genomics. *Nat. Rev. Genet.* **3**, 210–219 (2002).

7. Zeng, A. *et al.* Prospectively Isolated Tetraspanin + Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* **173**, 1593-1608.e20 (2018).

8. Nishimura, O. *et al.* Unusually large number of mutations in asexually reproducing clonal planarian Dugesia japonica. *PLoS One* **10**, 1–23 (2015).

9. Otto, S. P. & Orive, M. E. Evolutionary consequences of mutation and selection within an individual. *Genetics* **141**, 1173–1187 (1995).

10. Otto, S. P. & Hastings, I. M. Mutation and selection within the individual. *Genetica* **102**–**103**, 507–524 (1998).

11. Neher, R. A. & Shraiman, B. I. Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics* **188**, 975–996 (2011).

12. Waples, R. S. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**, 379–91 (1989).

13. Jorde, P. E. & Ryman, N. Temporal allele frequency change and estimation of effective size in populations with overlapping generations. *Genetics* **139**, 1077–1090 (1995).

14. Berthier, P. *et al.* Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics* **160**, 741–51 (2002).

15. Waples, R. S. & Yokota, M. Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175**, 219–233 (2007).

16. Lobo, D., Beane, W. S. & Levin, M. Modeling planarian regeneration: A primer for reverse-engineering the worm. *PLoS Comput. Biol.* **8**, (2012).

17. Baguñà, J., Saló, E. & Auladell, C. Regeneration and pattern formation in planarians \nIII. Evidence that neoblasts are totipotent stem cells and the source of blastema cells. *Development* **107**, 77–86 (1989).

18. Elliott, S. A. & Sánchez Alvarado, A. The history and enduring contributions of planarians to the study of animal regeneration. *Wiley Interdiscip. Rev. Dev. Biol.* **2**, 301–326 (2013).

19. Montgomery, J. R. & Coward, S. T. On the Minimal Size of a Planarian Capable of Regeneration. *Trans. Am. Microsc. Soc.* **93**, 386–391 (1974).

20. Pollak, E. A New Method for Estimating the Effective Population Size from Allele Frequency Changes. *Genetics* **104**, 531–548 (1983).

21. Pellettieri, J. Regenerative tissue remodeling in planarians – The mysteries of morphallaxis. *Semin. Cell Dev. Biol.* **87**, 13–21 (2019).

22. Campbell, P. J. & Martinocorena, I. Somatic mutation in cancer and normal cells. *Science (80-.).* **349**, 1483–1489 (2015).

23. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).

24. Sniegowski, P. D., Gerrish, P. J., Johnson, T. & Shaver, A. The evolution of mutation rate: separating causes from consequences. *BioEssays* **22**, 1057–1066 (2000).

25. Andre, J. B. & Godelle, B. The evolution of mutation rate in finite asexual populations. *Genetics* **172**, 611–626 (2006).

26. Dunham, J. P. & Friesen, M. L. A cost-effective method for high-throughput construction of

Illumina sequencing libraries. *Cold Spring Harb. Protoc.* **9**, 820–834 (2013).

27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

28. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* **270**, 313–21 (2003).

29. Kofler, R., Langmuller, A. M., Nouhaud, P., Otte, K. A. & Schlotterer, C. Suitability of Different Mapping Algorithms for Genome-wide Polymorphism Scans with Pool-Seq Data. *G3&amp;#58; Genes|Genomes|Genetics* **6**, 3507–3515 (2016).

30. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. 1–9 (2012).

31. Waples, R. S. Temporal Variation in Allele Frequencies: Testing the Right Hypothesis. *Evolution (N. Y).* **43**, 1236–1251 (1989).

Table 1. Estimates of stem cell N_e from the II-XII and II-XIV generation intervals and according to three scenarios of cell immortality, and slow and fast tissue turnover ($g = 0.5, 1\ and\ 5$, respectively). N_e calculations were based on 370 SNPs selected for having AAF 0.1-0.9 at the start of the experiment (generation II) and being covered 10-60X in all merged-rep samples. Somatic mutation rate was calculated using average nucleotide diversity $\bar{\pi} = 0.00398$ and assuming $\mu_{som} = \frac{\bar{\pi}}{2N_{e,sc}}$ under neutrality. g: ratio of cellular / organismal generations, $\bar{F}_k$: arithmetic average of F_k over loci, S_i: harmonic mean of sequencing coverage of loci at the initial sampling point (generation II), S_j: harmonic mean of sequencing coverage of loci at the final sampling point (generation XII or XIV as indicated in the first column), $N_{e,sc}$: effective size of stem cell population, μ_{som}: somatic mutation rate.

Generational interval	# organismal generations elapsed	g	# cellular generations elapsed	$\bar{F}_k$	S_i, S_j	$N_{e,sc}$	μ_{som}
II-XII	10	0.5	5	0.1433	31.25, 21.73	71.10	2.7989E-05
II-XIV	12	0.5	6	0.1378	31.25, 15.14	138.49	1.4369E-05
II-XII	10	1	10	0.1433	31.25, 21.73	141.70	1.4044E-05
II-XIV	12	1	12	0.1378	31.25, 15.14	276.48	7.1976E-06
II-XII	10	5	50	0.1433	31.25, 21.73	706.52	2.8166E-06
II-XIV	12	5	60	0.1378	31.25, 15.14	1380.39	1.4416E-06
II-XII	10	10	100	0.1433	31.25, 21.73	1412.54	1.4088E-06
II-XIV	12	10	120	0.1378	31.25, 15.14	2760.27	7.2094E-07

Table S1. Accession number of Tricladida COI barcode sequences used for taxonomical identification of specimens and the number of reads from two representative samples from generations 6 and 10 mapping to each barcode.

Accession	Species	Sample 1 (gen. 6)	Sample 2 (gen. 10)
GBPL4280-15	*Girardia sp. AW2014*	411	561
GBPL1167-09	*Girardia tigrina*	298	504
GBPL2853-13	*Obama carinata*	12	57
GBPL2870-13	*Choeradoplana bocaina*	5	23
GBPL2825-13	*Pasipha tapetilla*	3	13
GBPL2856-13	*Notogynaphallia plumbea*	1	9
GBPL2926-13	*Luteostriata graffi*	0	5
GBPL2839-13	*Luteostriata ceciliae*	1	3
GBPL1552-13	*Obama ladislavii*	1	2
GBMAB2254-15	*Platydemus manokwari*	1	2
GBPL2921-13	*Endeavouria septemlineata*	0	2
GBPL2834-13	*Choeradoplana gladysmariae*	0	1
GBPL2886-13	*Issoca jandaia*	0	1
GBPL2907-13	*Issoca rezendei*	0	1
GBPL3444-14	*Luteostriata abundans*	0	1
GBPL2861-13	*Notogynaphallia parca*	0	1
GBPL3196-14	*Cephaloflexa bergi*	1	0
GBPL1559-13	*Geoplana quagga*	1	0
GBPL2859-13	*Pasipha chimbeva*	1	0
GBPL1173-09	*Artioposthia testacea*	0	0
GBPL1172-09	*Bipalium adventitium*	0	0
GBPL2841-13	*Cephaloflexa araucariana*	0	0
GBPL2833-13	*Choeradoplana albonigra*	0	0
GBPL2858-13	*Choeradoplana banga*	0	0
GBPL1551-13	*Choeradoplana iheringi*	0	0
GBPL2837-13	*Cratera crioula*	0	0
GBPL2903-13	*Cratera pseudovaginuloides*	0	0
GBPL2901-13	*Cratera tamoia*	0	0
GBPL1169-09	*Dendrocoelum lacteum*	0	0
GBPL1458-13	*Dugesia aethiopica*	0	0
GBPL2668-13	*Dugesia ariadnae*	0	0
GBPL1478-13	*Dugesia benazzii*	0	0
GBPL2669-13	*Dugesia cretica*	0	0
GBPL1472-13	*Dugesia etrusca*	0	0
GBPL1469-13	*Dugesia gonocephala*	0	0
GBPL1468-13	*Dugesia hepta*	0	0
GBPL1467-13	*Dugesia ilvana*	0	0
GBMTG2496-16	*Dugesia japonica*	0	0

GBPL1464-13	*Dugesia liguriensis*	0	0
GBPL1463-13	*Dugesia notogaea*	0	0
GBPL1443-13	*Dugesia sicula*	0	0
GBPL1460-13	*Dugesia subtentaculata*	0	0
GBPL3422-14	*Enterosyringa pseudorhynchodemus*	0	0
GBPL2927-13	*Geobia subterranea*	0	0
GBPL2830-13	*Geoplana chita*	0	0
GBPL2865-13	*Geoplana goetschi*	0	0
GBPL2894-13	*Geoplana hina*	0	0
GBPL2850-13	*Geoplana pulchella*	0	0
GBPL2893-13	*Geoplana vaginuloides*	0	0
GBPL1071-09	*Girardia anderlani*	0	0
GBPL2828-13	*Gusana*	0	0
GBPL1546-13	*Imbira guaiana*	0	0
GBPL1549-13	*Imbira marcusi*	0	0
GBPL2847-13	*Luteostriata ernesti*	0	0
GBPL2849-13	*Luteostriata muelleri*	0	0
GBPL2902-13	*Matuxia tuxaua*	0	0
GBSP4320-12	*Microplana robusta*	0	0
GBSP4311-12	*Microplana terrestris*	0	0
GBPL2898-13	*Notogynaphallia sexstriata*	0	0
GBPL1547-13	*Obama burmeisteri*	0	0
GBPL2842-13	*Obama josefi*	0	0
GBPL2826-13	*Paraba franciscana*	0	0
GBPL2860-13	*Paraba multicolor*	0	0
GBPL2869-13	*Paraba phocaica*	0	0
GBMAB314-15	*Parakontikia ventrolineata*	0	0
GBPL2889-13	*Pasipha pasipha*	0	0
GBPL2891-13	*Pasipha pinima*	0	0
GBPL2913-13	*Pasipha rosea*	0	0
TERIN001-17	*Rhynchodemus sylvaticus*	0	0
GBMTG5013-16	*Schmidtea mediterranea*	0	0
GBPL1435-13	*Schmidtea polychroa*	0	0
GBPL2827-13	*Supramontana irritata*	0	0

Table S2. Subsampled-and-pooled reads mapping to the four Dugesiidae genomes

	NCBI WGS project	Local	End-to-end
G. tigrina	MQRK01	21'728	17'439
S. mediterranea CIW4	AUVC01	15	10
S. mediterranea S2	NNSW01	14	6
D. japonica	MQRL01	10	2
Total mapped reads		21'767	17'457
Total reads		24'000	24'000

Table S3. Details of quality control and mapping statistics of trimmed and deduplicated fastq files. The R1 fastq file from sequencing lane 8 of the generation 14 (sample XIV_L008) was damaged during a server transfer and could not be recovered. This sample was pre-processed like others but as a single-end sequenced library. The Raw fastqc files had varying levels of low quality read segments (especially towards the 3' end and more in the reverse read or 2nd mate files) and PCR duplicates. They did not contain adapter or any other over-represented sequences (other than N homopolymers in some cases). Across the 12 paired-end samples, the fractions of proper pairs after processing by Trimmomatic and Clumpify were 87-95% and 67-85%, respectively. Clumpify was very high-memory-consuming and worked for all samples only if the -Xmx option were set to 230g or higher. Min. deduplication (%): "percent of sequences remaining if deduplicated" from the output of FastQC after running Clumpify. This is a minimum estimate because it is based on read sequence identity inferred separately from forward and reverse files while Clumpify identifies duplication based on both mates of a pair. Mapping rate: percentage of trimmed and deduplicated read pairs that were mapped to the *G. tigrina* reference genome by *bwa mem*.

Sample	Raw pairs	Trimmomatic output pairs	Clumpify output pairs	Min. deduplication (%)	Mapping rate (proper pairs) (%)
II_L003	179'975'602	171'531'579	152'492'295	84.93; 77.74	96.22 (73.99)
II_L005	172'108'843	163'149'585	144'494'228	85.36; 77.76	96.23 (74.11)
II_L006	178'680'161	169'213'772	150'056'408	85.93; 78.82	96.20 (74.16)
VI_L002	200'302'395	189'399'008	149'305'526	84.83; 78.15	93.90 (70.94)
VI_L003	196'579'360	171'891'056	141'505'349	81.85; 80.74	92.81 (68.76)
VIII_L005	191'266'046	166'653'906	130'689'674	83.56; 83.9	93.19 (68.04)
VIII_L006	191'085'892	166'682'020	130'303'177	82.45; 80.27	93.47 (68.48)
X_L007	186'274'991	167'125'904	127'102'268	82.44; 80.46	92.81 (69.80)
X_L008	188'302'804	168'925'128	128'794'306	82.71; 81.52	92.82 (69.77)
XII_L005	184'648'701	165'342'241	130'904'629	81.48; 76.55	92.66 (69.01)
XII_L006	192'481'056	170'964'293	129'036'162	79.8; 80.61	92.17 (68.30)
XIV_L007	197'049'701	183'302'121	152'225'064	82.87; 82.2	92.64 (68.56)
XIV_L008*	190'083'700	169'937'444	110'184'051	92.09	84.80 (N/A)

* Only the reserve file was available from this sample. All numbers presented for this row represent single reads not pairs.

Table S4. Evaluating the goodness-of-fit of AFS with neutrality (drift) vs. strong linked selection (draft). Positions with derived allele frequency $\nu: 0.01 - 0.9$ were included to avoid extreme outlier effects and fixed substitutions. Variables $\nu.inv = \nu^{-1}$ and $\nu.inv.sq = \nu^{-2}$ were created and the linear fit of $f(\nu)$ against them was examined. The simple $f(\nu) \sim \nu$ linear model was tested as a known bad fit (negative control).

Model	Regression coefficient	Significance	Goodness-of-fit
$f(\nu) \sim \nu$	$\beta = -2.4616$	$p = 2 \times 10^{-16}$	$R^2 = 0.1604$
$f(\nu) \sim \nu^{-1}$	$\beta = 0.1574$	$p = 2 \times 10^{-16}$	$R^2 = 0.9375$
$f(\nu) \sim \nu^{-2}$	$\beta = 0.00211$	$p = 2 \times 10^{-16}$	$R^2 = 0.832$

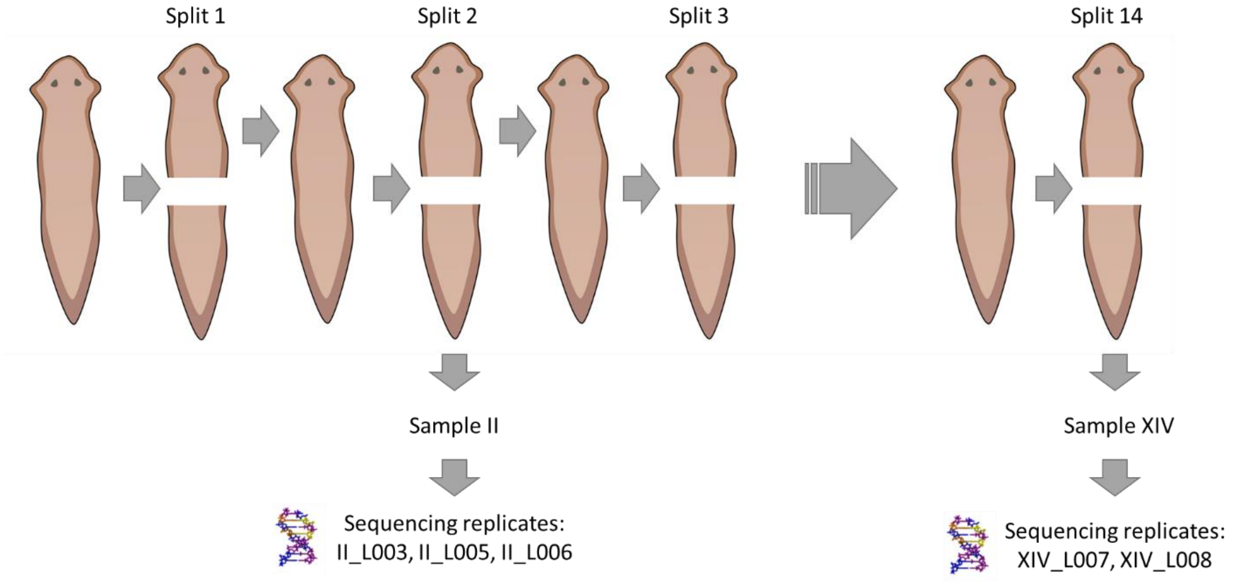

Figure 1. Experimental design of the project. Sample names have two parts: the first part is the split after which the biological sample was derived, i.e., organismal generation; the second part is the lane on the HiSeq machine where that replicate was sequenced. For example, sample II_L003 or II.3 was derived after split 2 (generation 2) and sequenced on lane 3. Sample XIV.8 was disqualified due to the loss of its forward sequencing fastq file effectively making sample XII (generation 12) the latest generation that was sequenced in replicates (XII.5 and XII.6). All paired-end sequenced libraries were used for taxonomic identification. Population genetic analyses mainly involved sample II (II.3, II.5, II.6) and XII (XII.5, XII.6).

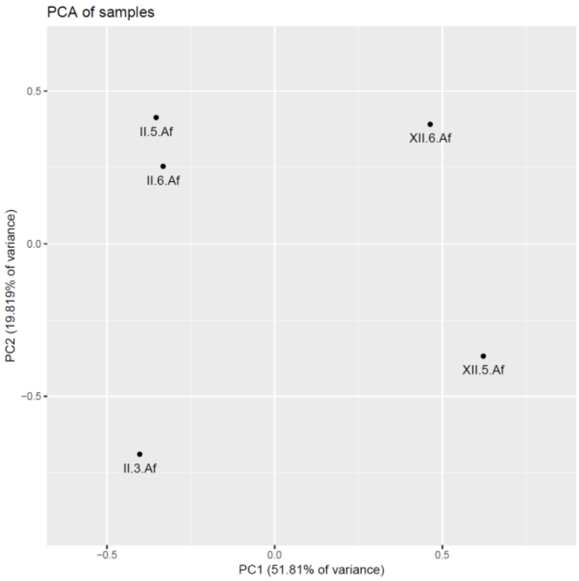

Figure S1. PCA of allele frequencies of replicates from samples II and XII. PC1 which explains ~52% of the variance (and more than twice as much as PC2) clearly separates II.3, II.5 and II.6 from XII.5 and XII.6. This confirms that variation due to drift is essentially larger than technical variability.

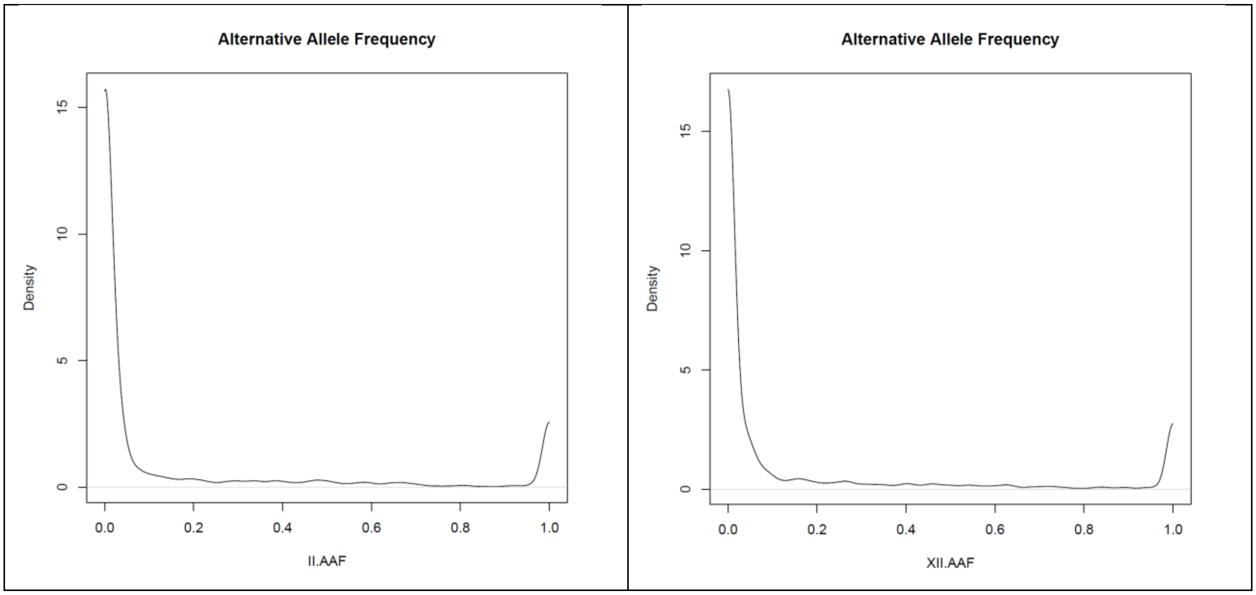

Figure S2. Allele frequency spectra of merged-rep samples II and XII. Positions with $f_{alt} = 1$ would not be normally included in the AFS because they are not polymorphic. They are recorded here because we ran FreeBayes with the --use-reference-allele option. They provide a visual comparison of divergence (fixed substitutions) from the reference versus segregating variation. AAF: alternative allele frequency.

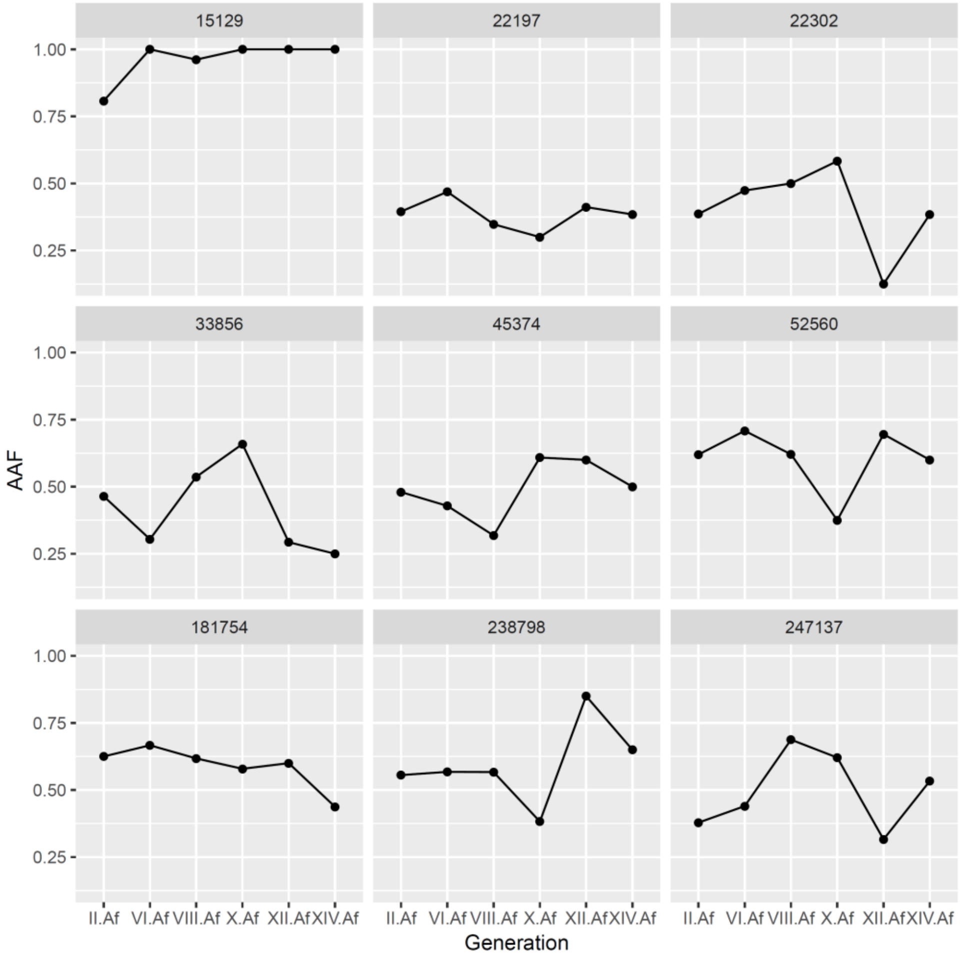

Figure S3. Transgenerational fluctuation of AAF (alternative allele frequency) in 9 randomly selected SNPs with AAF 0.1-0.9 at generation II. The same criterion was applied to include SNPs in calculation of $N_{e,sc}$. All positions were covered 10-60X in all samples. The observed changes in AAF are due to somatic drift between generations + sampling variance at each generation. Number on top of panels: position on scaffold MQRK01218062.1.

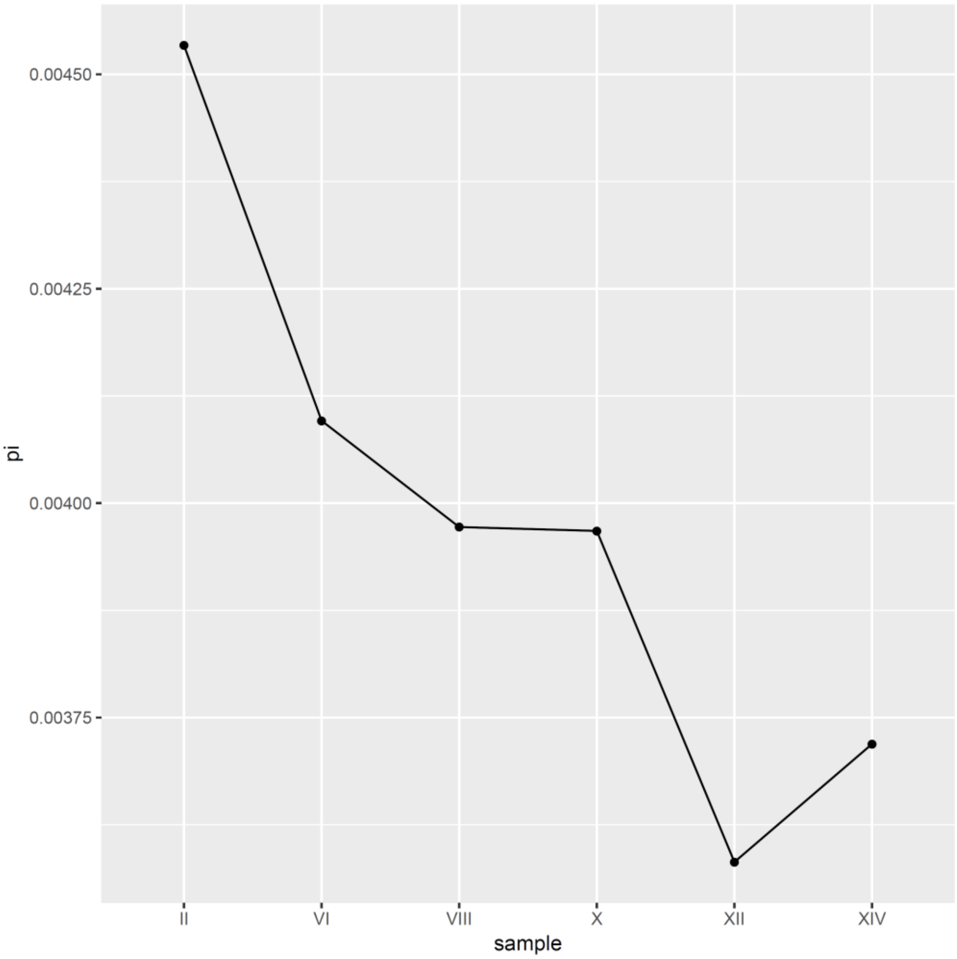

Figure S4. Decreasing trend of nucleotide diversity over generations.