The impact of genetic surfing on neutral genomic diversity

Flávia Schlichta[1,2,†], Stephan Peischl[2,3] and Laurent Excoffier[1,2,†]

[1] Computational and Molecular Population Genetics lab, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[2] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[3] Interfaculty Bioinformatics Unit, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[†] Corresponding authors: flavia.schlichta@unibe.ch; laurent.excoffier@unibe.ch

Keywords: Range expansions, genetic surfing, genomic diversity, genome scan

ORCID:

F. Schlichta: 0000-0002-8845-3623

S. Peischl: 0000-0002-0474-6104

L. Excoffier: 0000-0002-7507-6494

Emails:

F. Schlichta: flavia.schlichta@unibe.ch

S. Peischl: stephan.peischl@bioinformatics.unibe.ch

L. Excoffier: laurent.excoffier@unibe.ch

Abstract

Range expansions have been common in the history of most species. Serial founder effects and subsequent population growth at expansion fronts typically lead to loss of genomic diversity along the expansion axis. A frequent consequence is the phenomenon of “gene surfing”, where variants located near the expanding front can reach high frequencies or even fix in newly colonized territories. Although gene surfing events have been characterized thoroughly for a specific locus, their effects on linked genomic regions and on the overall patterns of genomic diversity have been little investigated. In this study, we simulated the evolution of whole genomes during several types of 1D and 2D range expansions differing by the extent of migration, founder events and recombination rates. We focused on the characterization of local dips of diversity, or “troughs”, taken as a proxy for surfing events. We find that, for a given recombination rate, once we consider the amount of diversity lost since the beginning of the expansion, it is possible to predict the initial evolution of trough density and their average width irrespectively of the expansion condition. Furthermore, when recombination rates vary across the genome, we find that troughs are over-represented in regions of low recombination. Therefore, range expansions can leave local and global genomic signatures often interpreted as evidence of past selective events. Given the generality of our results, they could be used as a null model for species having gone through recent expansions, and thus be helpful to correctly interpret many evolutionary biology studies.

Introduction

Range expansions are a ubiquitous phenomenon affecting the demographic history of most species, as they often happen during the colonization (or invasion) of new habitats or because of contractions and re-expansions caused by climate fluctuations. A prominent example for the latter are the genetic patterns left by the quaternary glacial cycles that deeply affected past biodiversity (Hewitt 2000). In addition, our own species history is heavily colored by range expansions: when humans expanded out of Africa (e.g. Handley et al. 2007; Henn et al. 2016), or when early farmers from the Near East colonized Europe during the Neolithic period (e.g. Hofmanová et al. 2016). Understanding the genetic effects of range expansions on a genomic level is thus critical to answer questions related to differentiation, speciation, and adaptation. Range expansions are often characterized by serial founder effects (Slatkin and Excoffier 2012), where a few individuals (the founders) leave a source population to occupy a new adjacent area, grow to carrying capacity and send further founders to colonize a new deme, a three-step process that is repeated throughout the range expansion. During a range expansion, these serial founder effects reduce genetic diversity along the expansion axis (Austerlitz et al. 1997), while migration and population growth mitigate this loss (Excoffier 2004; Klopfstein et al. 2006). Previous works have shown that range expansions lead to gene surfing (Edmonds et al. 2004; Klopfstein et al. 2006; Hallatschek and Nelson 2008; Hallatschek and Nelson 2010), that is when variants at the front of an expansion spread with the spatial colonization wave and go to fixation. Further studies also demonstrated that gene surfing can happen for variants under selection (Travis et al. 2007) and that even negatively selected variants can accumulate preferentially at the expansion front, leading to an "expansion load" (Peischl et al. 2013; Peischl et al. 2015; Peischl and Excoffier 2015). Range expansions also influence patterns of introgression and hybridization, namely enabling asymmetric introgression between species (Currat et al. 2008), an important factor to account for in speciation genomics. Finally, recent studies have shown that both positive selection and population contractions can lead to similar "dips of diversity" along the genome (hereafter called "troughs") and that these selective and neutral

sweeps can easily be confounded when analyzing single locus empirical data (Moinet et al. 2022).

Most previous range expansion studies have focused on the diversity loss occurring during range expansions either globally in the genome (Austerlitz et al. 1997; Peischl and Excoffier 2015) or at single loci, resulting in surfing events (Edmonds et al. 2004; Klopfstein et al. 2006; Hallatschek and Nelson 2008) or the formation of spatial regions of low diversity (or sectors, Hallatschek and Nelson 2008; Korolev et al. 2010). Here we aim to expand on this knowledge by examining the effects of range expansions on the overall patterns of genomic diversity and genomic regions linked to surfing events. We explore various demographic scenarios using forward-in-time simulations, recombination maps of varying complexity and perform genome scans to follow troughs dynamics during the expansion and analyze how diversity loss develops over time.

Results

Identification of troughs

Figure 1 - Genome diversity scans and through definition*. The three panels show how genomic diversity is lost and how troughs of diversity form during a 1D range expansion. Note that only a small section of the genome is shown here (1Mb). (A) Diversity scan at the beginning of the expansion, soon after the colonization of the first deme. (B) Diversity scan on the expansion edge deme, in the same genomic region, after the colonization of the 21st deme. (C) Same as A and B, but after the colonization of 30 demes. The solid blue line and the percentage on its right shows the proportion of diversity remaining relative to core diversity (average computed only for this 1Mb region). The dashed red line indicates the threshold used to determine troughs (in this example, 10% of average core diversity). Troughs are then identified as genomic regions with diversity below the threshold, using different arbitrary colors to represent distinct troughs. Dots are troughs formed by only one sliding window (10kb).*

We identified dips of diversity, or “troughs”, from nucleotide diversity genome scans. Troughs were defined as a series of one or more windows with mean nucleotide

diversity below a given fraction of the core's mean genetic diversity (10%, unless specified otherwise, fig. 1). Using this low threshold for trough definition, we could identify regions of the genome that are close to segments that have been surfing and have reached or were close to reaching fixation in the population, while still tolerating some mutations to have occurred during the expansion. Note that although this threshold is to some degree arbitrary, we obtained the same qualitative results in trough dynamics when considering different thresholds (5% and 20%, see supplementary fig. S1).

Regions of low diversity are commonly generated during range expansions

To understand the effects of demographic conditions prevailing during range expansions on gene surfing events, we simulated various demographic scenarios, differing in spatial distribution, migration rate between adjacent demes and number of founders (see Table 1). The parameter combinations were chosen to understand the effects of different intensities of genetic drift. For instance, scenarios with fewer founders or less migration correspond to scenarios with more drift than the reference scenario. Further details of the simulations are specified in the supplementary table S1 and in the Material and Methods section. For the sake of clarity, we will focus on six 1D or 2D scenarios that vary in the number of founders. The results obtained under the remaining scenarios are described in the Supplementary Material.

Table 1 – Properties of simulated demographic scenarios: varying the number of founders and migration rate to change the intensity of drift during the expansion.

Spatial distribution	Front Depth[a]	Front Width[b]	Migration rate[c]	Uniform Recombination	Number of Founders	alias
1D	5	1	0.1	1e-8	20	1D reference
	5	1	0.1	1e-8	10	1D fewer founders
	5	1	0.1	1e-8	30	1D more founders
	5	1	0.05	1e-8	20	1D less migration
	5	1	0.2	1e-8	20	1D more migration
2D	3	5	0.1	1e-8	10	2D reference
	3	5	0.1	1e-8	4	2D fewer founders
	3	5	0.1	1e-8	20	2D more founders
	3	5	0.05	1e-8	10	2D less migration
	3	5	0.2	1e-8	10	2D more migration

[a] Depth of the simulated expanding wave front. It includes the edge deme and those in the back of the expanding wave.

[b] Number of demes on the very edge of the simulated expanding front.

[c] Migration rate between adjacent demes as specified in the SLIM 3 forward simulator. Further details can be found in the Material and Methods section.

For all these scenarios we have used a uniform recombination rate of 1e-8 per bp per generation (Dumont and Payseur 2008), a migration rate of 0.1 and a uniform mutation rate of 1.25e-8 (Kong et al. 2012).

The results from scenarios with varying migration rates can be found in the supplementary materials.

1D range expansions

We characterized regions of extremely low diversity (troughs) by examining their genomic density, their size, and the proportion of the genome within these regions. As shown in figures 1, 2A and supplementary figure S2A, genetic diversity is progressively lost during the expansion, but at rates that vary depending on the demographic conditions. As expected, scenarios with fewer founders or less migration (fig. 2A and

S2A) have higher rates of loss of genetic diversity during 1D expansions, whilst scenarios with more founders or more migration lead to a slower loss of diversity over time (fig. 2A and S2A). Consequently, the remaining diversity at the end of the expansions is considerably higher for the *1D more founders and 1D more migration* scenarios than for our reference model, and the *1D fewer founders* and *1D less migration* show the strongest final diversity loss. These observations are in line with theoretical treatments of 1D or 2D range expansions or serial founder effects showing a progressive reduction of heterozygosity during the expansion (Austerlitz et al. 1997; Hallatschek and Nelson 2008; Korolev et al. 2010; DeGiorgio et al. 2011; Slatkin and Excoffier 2012). The proportion of the genome within troughs follows trajectories that are very similar to those of overall diversity loss (fig. 2B and supplementary fig. S2B), suggesting that average levels of diversity during a range expansion strongly depend on trough formation.

Figure 2 - Dynamics of trough formation during a 1D expansion*. We show trough dynamics for 1D models, varying in number of founders in comparison to the reference model (1D reference, green). Exact simulation parameters can be found in table 1 and supplementary table S1. (A) Dynamics of relative genetic diversity loss during the expansion. (B) Proportion of the genome within troughs at different stages of the expansion. (C) Evolution of trough density (number of troughs per Mb). (D) Dynamics of trough mean size. Solid lines represent mean statistics among all replicates and the corresponding shaded areas encompass 95% of the observed values.*

The number of troughs per Mb (trough density, fig. 2C and supplementary fig. S2C) displays very different dynamics during an expansion: it first increases quickly and then slows down until reaching a maximum (16-17 troughs/Mb, in the first 10-27 demes of the expansion, depending on the scenario), and then decreases to eventually reach a plateau by the end of the simulations for the scenario with fewer founders (around 6 troughs/Mb, fig. 2C). Contrastingly, the trough density for the *reference* scenario, the *more migration* and *more founders* scenarios do not reach a plateau in the simulated time, and keep declining until the end of the simulations. We posit that they would

probably also reach a plateau at a later stage if the expansion had been simulated for a longer time (supplementary fig. S2C and fig. 2C). A unimodal pattern is nevertheless observed in all scenarios and results from the dynamics of trough formation and size extension. The initial linear increase in trough density is due to the appearance of troughs caused by independent gene surfing events (e.g., fig.1A). As trough density increases, surfing events begin to occur next to one another, leading to a progressive merging of troughs (e.g., fig.1B). The maximum density corresponds to a point where the formation of new troughs is compensated by the merging of pre-existing ones. After this point, troughs will merge faster than new ones are (independently) created and trough density decreases until reaching a plateau. This plateau occurs because average genetic diversity is very low at the end of the expansion, close to the threshold used to define troughs, such that new mutations arising and increasing in frequency will locally increase genetic diversity and split preexisting large troughs into smaller ones. This plateau is the signature of a new mutation-drift equilibrium. In line with trough density dynamics, trough average size increases continuously during range expansions, a phenomenon driven by trough merging (fig. 2D and supplementary fig. S2D). Towards the end of the expansion, the *1D fewer founders* model has equilibrium trough sizes of 185kb on average, while they are of 175kb on average in the *1D less migration* scenario. We note that the variability of the observed values (shown by the shaded areas in figure 2 and in supplementary figure S2) of trough density and trough size for these models are much larger than observed for the three other scenarios with less genetic drift.

In the reference scenario, the maximum trough density is reached after the colonization of 19 demes (95 generations after the start of the expansion), at which point ~52.5% of the initial diversity was lost (supplementary table S2). Initially, troughs are mostly of the size of the sliding window used to compute diversity (10 kb), and then slowly increase in size to reach an average of 23Kb at the maximum density of 17 troughs per Mb, at which point they represent 39% of the genome (supplementary table S2). When the number of founders (or migration rate) is smaller, drift is stronger and chromosomal segments go to fixation more quickly. Therefore, there is less time for recombination to erode troughs, leading to overall longer troughs (fig. 2D and supplementary figure S2D).

In this case, since there are fewer recombination events during trough formation, the variance in trough size is larger than in the reference scenario, explaining the larger span of the statistics reported in figure 2 and supplementary figure S2 for the *1D less founders* and *1D less migration* scenarios, respectively. Moreover, the number of troughs also increases faster than in the reference scenario, because the probability of fixation of any nucleotide is inversely proportional to the effective size of the population, which is smaller in these cases. In the scenarios with larger migration rates and more founders, loss of diversity and genetic drift are less severe than in the reference scenario, and they thus show trends opposite to the scenarios with lower number of founders or migration rate.

2D range expansions

The 2D scenarios behave qualitatively similarly to 1D scenarios, showing a gradual loss of genetic diversity over time (fig. 3A and supplementary figure S3A). Note, however, that diversity is lost at a slower pace than during 1D expansions (2D scenarios ran 3 times longer than 1D), which is a result of migration from adjacent demes at the expansion front, which partially restores diversity, since neighboring demes often lose diversity at different loci.

Figure 3 - Dynamics of trough formation during a 2D expansion*. This figure is analogous to figure 2, but for 2D expansions.*

Like in 1D scenarios, the proportion of the genome within troughs (fig. 3B and supplementary fig. S3B) follows the same pattern as the diversity loss over time. Likewise, the dynamics of trough density (fig. 3C and supplementary fig. S3C) was very similar to that of 1D scenarios: a fast increase in trough density, reaching a maximum density similar to the 1D scenarios (16-18 troughs/Mb), then declining and reaching a plateau towards the end of the expansion (except for the *2D more founders* scenario).This plateau is maintained by a mutation-migration-drift equilibrium, as the loss of troughs can be compensated by a local increase in genetic diversity that can

arise either by *de novo* mutations or by new variants introduced in the sampled deme by migration. Dynamics of the mean size of troughs was also similar to 1D scenarios, with a phase of continuous increase followed by a plateau (fig. 3D and supplementary fig. S3D), with the *2D fewer founders* scenario showing the largest troughs (~100kb) for the 2D expansions. Interestingly, the scenario with more founders presented the smallest troughs (maximum size of 37.5 kb reached after 1465 generations, see supplementary table S2), but with a trend of continuous increase in size, suggesting that there was not enough time to reach mutation-migration-drift equilibrium. This lack of equilibrium also explains the continuous density increase observed in this model, indicating that trough fusion is still not important enough to counteract the fission of troughs caused by migration from adjacent demes.

We studied the effects of front size in 2D expansions by performing additional simulations with a narrow (3 demes) or wide (10 demes) front (see supplementary table S3). The results are extremely similar to the reference 2D front width of 5 demes. However, expansions with a narrow front lose diversity faster, reaching maximum trough density (~17 troughs/Mb) earlier and with equilibrium-size troughs of ~104 kb (supplementary fig. S4A-D). These results are quite similar to those obtained with the *2D fewer founders* scenario. Additionally, the *2D wide* scenario behaved similarly to the *2D more founders scenario*, with a continuous increase in trough size and density, never reaching mutation-migration-drift-equilibrium in the simulated time (supplementary fig. S4C and S4D), as a wider front allows different chromosome segments to surf and fix on the edge of the expansion and indirect migration from non-neighboring demes will restore locally lost diversity.

The amount of diversity lost since the onset of the expansion predicts initial trough dynamics

The rate of diversity loss should be tightly linked to the effective population size at the expansion front and can thus be interpreted as a measure of the strength of genetic drift. We next investigate how well this measure of genetic drift explains the spatial dynamics of genomic pattern formation during range expansions. We thus examined

how trough statistics varied as a function of diversity loss relative to the initial diversity present in the core population, which can be considered as a cumulative drift measure, across the different scenarios we studied.

Interestingly all 1D and 2D scenarios show extremely similar trajectories for the three reported trough statistics, and only start to slightly diverge for large levels of diversity loss (~>55%). The proportion of the genome in troughs increases almost linearly with relative diversity loss (fig. 4A). Likewise, mean trough sizes and trough density evolve almost identically for all demographic scenarios until about 55% of initial genetic diversity has been lost (fig. 4B and C, respectively). Note that this latter value of diversity loss corresponds to the maximum trough density, which is comparable for almost all scenarios (see supplementary table S2), except for the *2D more founders* scenario, where density keeps increasing for large values of diversity loss. After this apex point, the dynamics of all scenarios begin to diverge and show considerable differences when diversity loss is larger than 70%. The time taken to reach the maximum trough density (and thus the divergent behavior) depends on the amount of genetic drift present in each model: models with more drift start to diverge earlier than other scenarios (see figures 2C and 3C, scenarios with fewer founders compared to the reference scenarios). Globally, we thus find that the similarities in trough dynamics observed between 1D and 2D scenarios when displayed as function of time or distance from the core (fig. 2 and 3), become an almost exact match when expressed as a function of diversity loss (fig. 4).

Figure 4 - Dynamics of trough formation as a function of diversity loss for 1D and 2D expansions*. (A) Proportion of the genome within troughs, (B) trough density and (C) average trough size. All metrics are shown for different levels of diversity loss in the edge deme compared to the source population. Interestingly, the initial dynamics of trough formation is shown to be similar across all demographic scenarios. After about 55% relative diversity loss, scenarios start to diverge.*

Troughs are more often observed in low recombination regions

We examined the effect of a heterogeneity of recombination rate on trough formation by doing simulations under the *1D reference* scenario with four different heterogeneous recombination maps, each having the same average recombination rate (as described in supplementary table S4 and illustrated in supplementary fig. S5). Overall, we find that trough dynamics on chromosomes with different heterogeneous recombination rates are nearly identical to each other (supplementary fig. S6) and very similar to those occurring on a chromosome with the same average uniform recombination rate (supplementary fig. S7). However, using a permutation test, we find that during most of the range expansion, but especially in the beginning, detected troughs show a lower average recombination rate than expected by chance (i.e., in troughs randomly distributed over the chromosome) (see fig. 5A and 5B for 1D and 2D scenarios simulated under the "*Complex 100k*" recombination map). In older expansions, when diversity is lower and troughs are becoming very large, the average recombination rate in troughs becomes close to what is expected by chance. Like other trough statistics (fig. 4), the probability to observe a lower recombination rate in troughs behaves very similarly across all 1D and 2D scenarios when plotted against the relative diversity loss (fig. 5C). Interestingly, we find that the lower recombination rate in troughs is due to a small excess of troughs in lower recombination regions (2% or less on average, fig. 6). We also observe similar result for all other heterogeneous recombination maps, with troughs occurring more often in regions of lower average recombination rate than expected by chance (supplementary fig. S8). This pattern is again caused by an excess of troughs in low recombining regions and a deficit of troughs in high recombining regions, albeit to different degrees depending on the recombination landscape (see supplementary fig. S9).

In figure 5C, we see a dip in the probability of observing a lower recombination rate in troughs for all simulated expansions when diversity loss is around 0.2. We have no specific explanation for this observation, but we suspect it is due to a complex interaction between trough size, recombination map, and the width of the sliding window used to detect troughs. Note that this "dip" is not present for heterogeneous maps with

larger chunks of recombination categories (e.g., in Simple and Complex 1Mb scenarios, supplementary fig. S8C).

***Figure 5 – Low recombination rate within troughs**. We report here the probability that the average recombination rate in troughs is lower than expected by chance, for the 1D and 2D scenarios (panels in rows A and B, respectively). The shaded areas encompass 80% of the observed values (between the 10th and 90th percentiles of the simulated values) and the solid lines are the average value over all simulations. In the first two rows, this probability is shown on the expansion front as a function of time since the beginning of the expansion. In panel C, we report this probability as a function of the proportion of diversity lost on the front since the beginning of the expansion. Values above 0.5 indicate that observed troughs are preferentially observed in regions of low recombination rates, a signal commonly seen for selection. Solid lines in pane C are the average of each demographic scenario and dashed lines are the 10th percentile. The average and the percentiles were smoothed using a rolling window approach, where the central window is smoothed out considering the 3 neighbouring values on each side.*

***Figure 6 – Excess of troughs in low recombination regions**. **We report here for three 1D scenarios the excess or deficit of regions of low, medium or high recombination in troughs, relative to their expected values.** Shaded areas encompass 80% of the simulated values (between the 10th and 90th percentiles of the simulations). The light gray dashed line indicates no deviation from the expected proportions. There is initially a larger proportion than expected of troughs in low recombination regions, and a smaller proportion of troughs in medium or high recombination rate regions. This pattern is valid for all scenarios, but it persists for a longer time in the* 1D more founders *scenario, where drift is less intense, and troughs are smaller.*

Robustness analysis

We checked that our results remained valid under different sampling regimes and trough definitions. For instance, trough dynamics is qualitatively similar if we sample individuals on the edge demes at a different time of the growth period (supplementary

fig. S10). Moreover, the number of sampled genomes has little influence on our recorded statistics (supplementary fig. S11), and sampling demes at different time intervals leads to essentially the same dynamics with smoother curves for longer sampling intervals (supplementary fig. S12). We also checked the importance of the number of replicates and show that with 200 simulations per scenario, as used throughout this work, trough dynamics is identical to that computed from 1000 simulations (supplementary fig. S13). The size of the windows in genome scans however has a clear effect on the absolute values of trough size and density, since the window size defines the minimum through size and our ability to detect changes in trough density (supplementary fig. S14). For the sake of computational efficiency in 2D simulations, we only reported simulations where the moving front was 3 demes deep, and we ignored demes further away from the edge of the expansion. However, simulations done with a 5-demes deep front lead to no difference in the resulting trough statistics (supplementary fig. S15), suggesting that simulating a 3-demes deep front is adequate for predicting changes in genomic diversity during 2D expansions. Finally, we studied the effect of the diversity threshold used for trough identification on our trough statistics (supplementary fig. S1). As expected, using a lower threshold (5%) leads to narrower troughs and a smaller proportion of the genome in troughs. The trough density dynamics is also different, with an initially smaller density than for the reference threshold of 10%, as the use of a lower level of diversity to define a trough implies a longer time necessary to reach this threshold. Correspondingly, the final trough density is larger than for higher thresholds (supplementary fig. S1), because fewer mutations and/or migrations are necessary to locally increase diversity within an existing trough such as to split large troughs into smaller ones. As expected, trough dynamics with a higher threshold (20%) follows trends opposite to those described for the 5% threshold (i.e., larger proportion of the genome within troughs, lower final density, and larger size). In summary, even though the absolute values of the trough statistics depend on their detection threshold, we found that the overall qualitative behavior does not (supplementary fig. S1).

Discussion

We used individual-based simulations to investigate the spatial patterns of diversity loss across genomes during range expansions. Our key result is that the initial dynamics of trough formation can be entirely predicted by the amount of diversity lost since the beginning of an expansion, and that this initial dynamics, which ends when maximum trough density is reached, is similar for all types of range expansions we have considered. We also investigated the genome-wide effects of gene surfing caused by range expansions. As expected, the overall level of diversity in edge demes decreases during an expansion, and our simulations suggest that this process is initially tightly linked to trough formation and the fixation of short chromosomal segments. Gene surfing events mainly occur at independent sites without much overlap during the initial phase of the expansion, and troughs are rapidly created, saturating the genome until a maximum trough density is achieved (fig. 2C-D and 3C-D). After this apex point, the merging of existing troughs is faster than the creation of new ones on a trough-saturated chromosome, leading to fewer but longer troughs. However, this decrease then slows down until an equilibrium between emergence of new troughs and the eradication of existing ones by mutation and migration is reached. The only exception observed is the *2D more founders* scenario (fig. 4B*)*, where this equilibrium state is never achieved and new troughs continuously emerge, while average trough size is constantly increasing.

Genetic drift on the front determines initial trough dynamics

Specific demographic conditions during the expansions affect the spatial and temporal dynamics of trough formation, as well as the speed at which genetic diversity decreases. In brief, the stronger the genetic drift prevailing on the front, the faster the fixation of genome segments, resulting in fewer, but larger troughs at the end of the expansion (fig. 2D and 3D). Therefore, stronger drift also causes the number of troughs to increase more rapidly compared to other scenarios, achieving maximum trough density quicker (fig. 2C and 3C).

Since genetic drift determines the rate of diversity loss observed in each scenario, it also determines genomic trough dynamics. When studied as a function of diversity loss, trough formation and extension is extremely similar across all envisioned 1D and 2D scenarios until about half (~55%) of the initial diversity has been lost (fig. 4). This is interesting because it shows that the exact demographic conditions of an expansion do not need to be known to be able to predict trough patterns across the genome if the proportion of diversity loss is known. In practice, a good proxy of the initial level of diversity could be populations in the core of the range, close to the source of an expansion. Our results should thus be applicable to a variety of organisms that went through a recent expansion, and thus provide a reasonable null model of their expected neutral patterns of genomic diversity.

The fact that the dynamics of trough formation diverges only after some apex trough density is reached is puzzling. A possible explanation for this change in behavior is that the initial phase is dominated by the emergence of independent troughs due to the fixation of short chromosome segments, and that, for a given average recombination rate, this process only depends on the effective population size on the front. The overall rate of loss of genetic diversity on the edge of the expansion also depends on the effective population size on the front, such that expressing trough statistics as a function of diversity loss makes their dependence on effective size disappear. However, whereas trough occurrence only depends on population effective size, trough merging, which governs the second phase of the trough dynamics, depends on the extent of migration bringing back diversity and on the variation in trough size, such that the trough merging dynamics differs among scenarios.

Excess of troughs in low recombination regions

We have seen that trough properties and dynamics depend on the speed at which chromosomal segments go to fixation (fig. 2 and 3), which conditions the number of recombination events occurring during trough formation. Longer fixation times imply more recombination events and therefore smaller troughs. The impact of recombination rate on trough properties is thus easy to predict: lower recombination rates promote the

formation of less numerous but larger troughs. This is confirmed in simulations performed on chromosomes having homogeneous but different recombination rates (supplementary fig. S16). However, recombination rates usually vary along chromosomes (McVean et al. 2004), and it is important to understand how heterogeneity of recombination rates along a chromosome will affect trough formation dynamics and genomic distribution. We have therefore simulated expansions and the evolution of chromosomes with different heterogeneous recombination maps (supplementary table S4 and fig. S5). In that case, we find that troughs are found more often in regions of low recombination rate for both 1D (fig. 5A) and 2D (fig. 5B) range expansions. Moreover, this pattern is similar in all demographic scenarios when scaled to relative diversity loss (fig. 5C). Interestingly, the significantly lower recombination rate found in troughs is only due to a relatively small excess of low-recombination segments within troughs (<2%, fig. 6). This excess is not due to a higher probability of gene surfing events occurring in lower recombination regions as the fixation probability of a given chromosome segment depends on the effective population size, which should be similar over the whole genome, irrespective of the underlying recombination rate in absence of selection, as simulated here. Thus, this excess is not due to a higher number of troughs in regions of low recombination, but it should be due to a higher probability of trough detection in low recombination regions, since troughs will extend more (become wider) if occurring within or in the vicinity of regions of lower recombination.

Disentangling the effects of demography from selection on genomic diversity data has been an important pursuit in evolutionary biology (Li et al. 2012; Bank et al. 2014; Lohmueller 2014; Charlesworth and Jensen 2021). In this respect, dips of genetic diversity are often considered indicative of recent selective sweeps (Pavlidis et al. 2008; Pavlidis et al. 2010; Stephan 2019) and should accumulate in regions of low recombination (Begun and Aquadro 1992). Indeed, many candidate loci for positive selection have been identified in regions of low recombination in humans (O'Reilly et al. 2008). The fact that troughs caused by neutral sweeps also tend to span regions of low recombination is important, since it shows that a purely neutral process can reproduce a

signal previously attributed to selection at the whole-genome level and not only at single-locus level (Koropoulis et al. 2020; Moinet et al. 2022).

Limitations of the study

We have shown that several signals attributed to positive selection commonly arise during range expansions. However, this is only a first step towards the full whole-genome characterization and identification of gene surfing events, as several issues remain to be addressed. Our results are indeed limited to what happens during an expansion process, and we have not yet studied how genetic diversity recovery after the end of an expansion will affect trough properties. Such inquiry should also help understand for how long the footprints of past expansions could still be detected after they end. Another important factor to be studied would be the interaction between gene surfing events and background selection (BGS), since the extent of their effects depends on local recombination rates, and they can both mimic the effect of selective sweeps (Jensen 2014; Charlesworth and Jensen 2021), even though this similarity has been challenged recently (Schrider 2020).

Future applications

It would be tempting to apply our findings to real world data, by examining the distribution of troughs over the genome of some species having gone through some recent range expansion, but a precise knowledge of the effect of BGS and population recovery on neutral sweeps may be required before such attempt. In any case, since selected sweeps and neutral sweeps could lead to very similar dips of diversity at individual loci (Moinet et al. 2022), it seems that the distinction between selection and demography would be better achieved by examining overall genomic patterns rather than looking at individual windows (Schrider and Kern 2018). In this respect, an interesting goal would be to infer the likelihood of a given distribution of trough density and size predicted from a given distribution of fitness effect previously inferred from patterns of genomic diversity (Kim et al. 2017; Tataru and Bataillon 2020) and compare it to that expected under a range expansion scenario.

Material and Methods

Simulated expansion dynamics

We simulated range expansions using the forward-in-time SLiM 3 simulation software (Haller and Messer 2019), based on a standard Wright-Fisher model. Generations are discrete and non-overlapping, individuals are diploid, monoecious and reproduce through biparental random mating. The diploid genome of each individual consists of two homologous autosomes of 100 Mb, with a constant and uniform mutation rate (1.25e-8 per bp per generation, Table S1). This mutation rate is in the range of the average genome-wide rates estimated in humans (Kong et al. 2012; Narasimhan et al. 2017) and it has been used in several other simulation studies (e.g. Speidel et al. 2019; Almarri et al. 2021; Excoffier et al. 2021). In our reference scenario, we also assumed a uniform and constant recombination rate of 1e-8 per bp per generation, equivalent to 1 cM/Mb, which is close to what is estimated for humans (1.144 cM/Mb, Dumont and Payseur 2008). All mutations are neutral, and in case of multiple mutations per site, only the last mutation that occurs is kept at a given site (stacking policy = "l" in SLiM). A large core population (N=2500) is first created and let to evolve for 10N generations to reach drift-mutation equilibrium as a burn-in phase. An expansion is then initiated by sending a given number of emigrants (founders) to the closest empty deme from the core, thus founding a new deme, as in a standard stepping-stone model (supplementary fig. S17A). The population in the new edge deme grows exponentially for five generations, until it reaches carrying capacity (N_{MAX}). At this stage, the edge deme sends founder individuals to the next neighboring empty deme, increasing the size of the colonized world.

We studied two different spatial configurations. First, we modelled an expansion along a one-dimensional array of demes connected by migration as in a stepping-stone model (hereafter simply referred to as "1D" model). Second, we modelled an expansion across a two-dimensional array of demes arranged in a lattice with migration to and from the four nearest neighbors (supplementary fig. S17B). To eliminate border effects the upper

demes are connected to the bottom demes such that the expansion occurs on the surface of a cylinder (torus). This model is hereafter referred to as the 2D model. In all simulated models, migration happens according to SLIM 3 standard model, in which the migration rate between two demes is used to define the proportion of offspring (i.e., the population of the next generation) which will have at least one parent originating from the other population. While SLIM 3 allows for asymmetric rates, migration rates between two demes were always simulated as symmetric.

To save computational time and memory, we only simulated a given number of connected demes away from the wavefront. Unless specified otherwise, the simulated depth of the 1D moving front was set to 5 demes, and the depth of the 2D moving front was set to 3 demes (see supplementary figure S15 for details showing that a 3-deme deep wavefront is sufficient). It implies that when a 1D expansion colonizes the 6th deme, the wavefront is detached from the core. Similarly, this detachment occurs after the fourth colonization event in the 2D model. We simulated an expansion over 100 demes for the 1D model and a longer expansion over 300 demes for the 2D model, because the dynamics of trough formation are generally slower in 2D than in 1D (see Results section).

Genome scan: identification of low diversity segments (troughs)

Throughout the expansion, we sampled ten randomly chosen individuals from the leading edge of the expansion every five generations. For 2D expansion, the sampled deme was in the middle of the front edge. Average nucleotide diversity was calculated along the genome in sliding windows of 10kb, with a 25% overlap between adjacent windows. We then characterized as troughs the regions of the genome below a certain threshold. This threshold was a fraction (5%, 10% or 20%) of the core population's average genetic diversity. Since troughs are strictly delimited by the threshold, troughs separated by a single window exceeding the threshold are considered as separate troughs, which yields a fine-scale picture of potentially surfing chromosome segments. We characterized the genomic dynamics of these troughs during expansions using three statistics: the trough density (computed as the number of troughs per Mb), the

average trough length, and the proportion of the genome within troughs (i.e., below the threshold).

Simulated scenarios

In order to better understand how some demographic parameters affect trough formation during range expansion, we simulated several demographic scenarios under different spatial configurations. We thus varied migration rate between adjacent demes (5%, 10% or 20%, Deshpande et al. 2009) and the number of founders (10, 20 or 30 individuals) for 1D models. For the 2D scenarios, we varied migration rates (5%,10% and 20%), the number of founders (4, 10 and 20) and different widths of the expansion front (3, 5 or 10 demes). These demographic parameters were studied with different recombination maps, producing 19 individual scenarios, each investigated with 200 replicates of the simulation of 100 Mb genomes (see all scenarios in table 1 and supplementary table S3 and S4).

Two different types of recombination maps were used: uniform and heterogeneous. In the uniform maps, recombination was homogeneous over the entire genome, with the values of 1e-9, 1e-8, or 1e-7 per bp, referred to here as “low”, “medium” and “high” recombination scenarios, respectively. The heterogeneous maps consisted of a mixture of these three recombination rate categories (see supplementary fig. S5 and table S4). The main heterogeneous map used in this study (named “Complex 100kb”) was made up of a thousand identical blocks of 100kb, where each consisted of a 30kb long segment of low recombination rate, followed by six 10kb segments of medium recombination rate separated by five 2kb segments of high recombination rate (see supplementary fig. S5 and table S4). These proportions of different recombination rate segments were chosen to resemble those occurring in the human genome, and the high recombination regions we use are equivalent in magnitude to the hotspots of recombination found in McVean et al. (2004). To detect possible biases in trough formation regarding local recombination rates in simulations with heterogeneous recombination, we performed a non-parametric permutation test described hereafter. At each generation we recorded the number and lengths of troughs and calculated the mean recombination rate within these troughs based on their recombination maps

shown in supplementary figure S5. We generated a null trough distribution by shuffling the observed troughs over the genome, maintaining the number and length of observed troughs, and then re-calculating the mean recombination rate within troughs. This procedure was repeated 100 times for each generation and each replicate, and we counted the number of times in which the average recombination rate within observed throughs was smaller than the value calculated after each permutation. This number was then divided by the total number of trials, thus obtaining the probability that troughs are more commonly found in regions of low recombination than what would be expected by chance.

Acknowledgements

FS was supported by a grant from the Swiss National Science Foundation (No 310030_188883) to LE.

Data availability statement

Scripts used to run simulations and perform data analyses will be publicly available as a git repository (https://github.com/CMPG/genomicSurfing)

References

Almarri MA, Haber M, Lootah RA, Hallast P, Al Turki S, Martin HC, Xue Y, Tyler-Smith C. 2021. The genomic history of the Middle East. *Cell* 184:4612-4625.e14.

Austerlitz F, Jung-Muller B, Godelle B, Gouyon P-H. 1997. Evolution of Coalescence Times, Genetic Diversity and Structure during Colonization. *Theoretical Population Biology* 51:148–164.

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. 2014. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics* 30:540–546.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* 356:519–520.

Charlesworth B, Jensen JD. 2021. Effects of Selection at Linked Sites on Patterns of Genetic Variability. *Annual Review of Ecology, Evolution, and Systematics* 52:177–197.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution*:1908–1920.

DeGiorgio M, Degnan JH, Rosenberg NA. 2011. Coalescence-Time Distributions in a Serial Founder Model of Human Evolutionary History. *Genetics* 189:579–593.

Deshpande O, Batzoglou S, Feldman MW, Luca Cavalli-Sforza L. 2009. A serial founder effect model for human settlement out of Africa. *Proceedings of the Royal Society B: Biological Sciences* 276:291–300.

Dumont BL, Payseur BA. 2008. Evolution of the Genomic Rate of Recombination in Mammals. *Evolution* 62:276–294.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences* 101:975–979.

Excoffier L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol* 13:853–864.

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*:btab468.

Hallatschek O, Nelson DR. 2008. Gene surfing in expanding populations. *Theoretical Population Biology* 73:158–170.

Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model.Hernandez R, editor. *Molecular Biology and Evolution* 36:632–637.

Handley LJL, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. *Trends in Genetics* 23:432–439.

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci USA* 113:E440–E449.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* 405:907–913.

Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-del-Molino D, van Dorp L, López S, Kousathanas A, Link V, et al. 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences* 113:6886–6891.

Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. *Nat Commun* 5:5281.

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics* 206:345–361.

Klopfstein S, Currat M, Excoffier L. 2006. The Fate of Mutations Surfing on the Wave of a Range Expansion. *Molecular Biology and Evolution* 23:482–490.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir Aslaug, Jonasdottir Adalbjorg, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488:471–475.

Korolev KS, Avlund M, Hallatschek O, Nelson DR. 2010. Genetic demixing and evolution in linear stepping stone models. *Rev. Mod. Phys.* 82:1691–1718.

Koropoulis A, Alachiotis N, Pavlidis P. 2020. Detecting Positive Selection in Populations Using Genetic Data. In: Dutheil JY, editor. Statistical Population Genomics. Methods in Molecular Biology. New York, NY: Springer US. p. 87–123. Available from: https://doi.org/10.1007/978-1-0716-0199-0_5

Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M. 2012. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology* 21:28–44.

Lohmueller KE. 2014. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLOS Genetics* 10:e1004379.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science* 304:581–584.

Moinet A, Schlichta F, Peischl S, Excoffier L. 2022. Strong neutral sweeps occurring during a population contraction. *Genetics*:iyac021.

Narasimhan VM, Rahbari R, Scally A, Wuster A, Mason D, Xue Y, Wright J, Trembath RC, Maher ER, van Heel DA, et al. 2017. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun* 8:303.

O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.* 18:1304–1313.

Pavlidis P, Hutter S, Stephan W. 2008. A population genomic approach to map recent positive selection in model species. *Molecular Ecology*:3585–3598.

Pavlidis P, Jensen JD, Stephan W. 2010. Searching for Footprints of Positive Selection in Whole-Genome SNP Data From Nonequilibrium Populations. *Genetics* 185:907–922.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. 2013. On the accumulation of deleterious mutations during range expansions. *Mol Ecol* 22:5972–5982.

Peischl S, Excoffier L. 2015. Expansion load: recessive mutations and the role of standing genetic variation. *Mol Ecol* 24:2084–2094.

Peischl S, Kirkpatrick M, Excoffier L. 2015. Expansion Load and the Evolutionary Dynamics of a Species Range. *The American Naturalist* 185:E81–E93.

Schrider DR. 2020. Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps. *Genetics* 216:499–519.

Schrider DR, Kern AD. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics* 34:301–312.

Slatkin M, Excoffier L. 2012. Serial Founder Effects During Range Expansion: A Spatial Analog of Genetic Drift. *Genetics* 191:171–181.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet* 51:1321–1329.

Stephan W. 2019. Selective Sweeps. *Genetics* 211:5–13.

Tataru P, Bataillon T. 2020. polyDFE: Inferring the Distribution of Fitness Effects and Properties of Beneficial Mutations from Polymorphism Data. *Methods Mol Biol* 2090:125–146.

Travis JMJ, Munkemuller T, Burton OJ, Best A, Dytham C, Johst K. 2007. Deleterious Mutations Can Surf to High Densities on the Wave Front of an Expanding Population. *Molecular Biology and Evolution* 24:2334–2343.

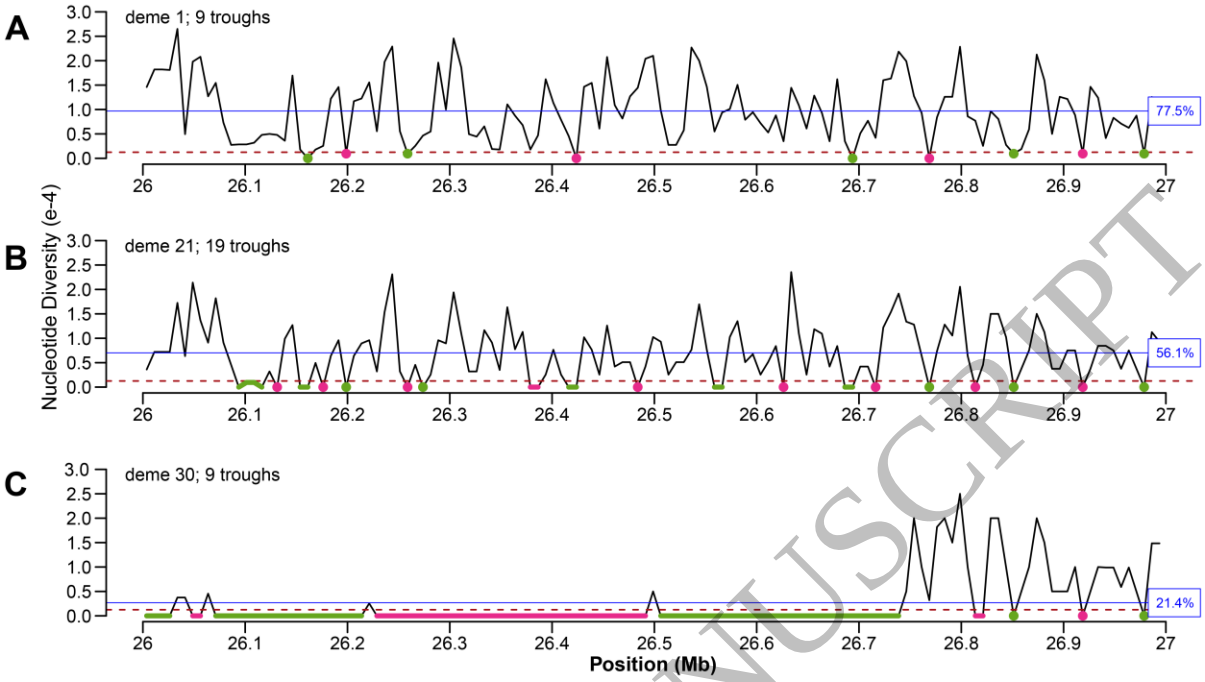

Figure 1
159x90 mm (.41 x DPI)

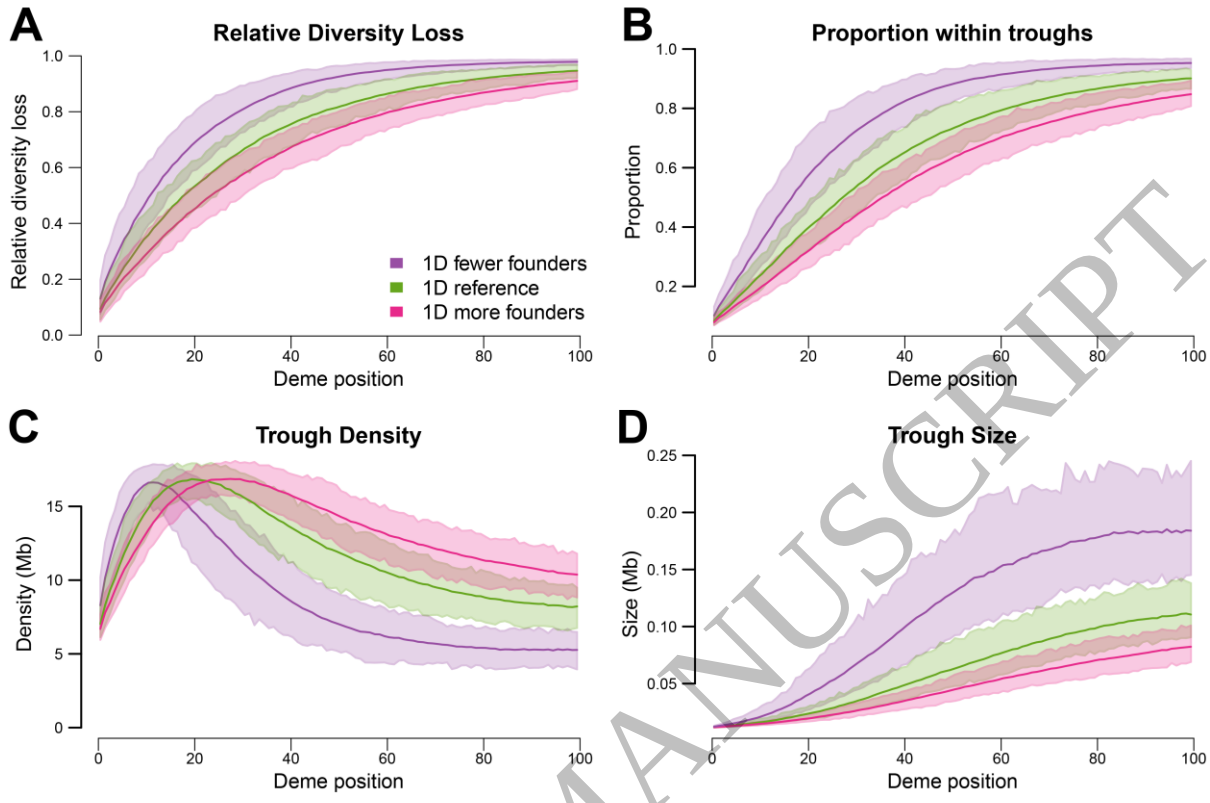

Figure 2
159x106 mm (.41 x DPI)

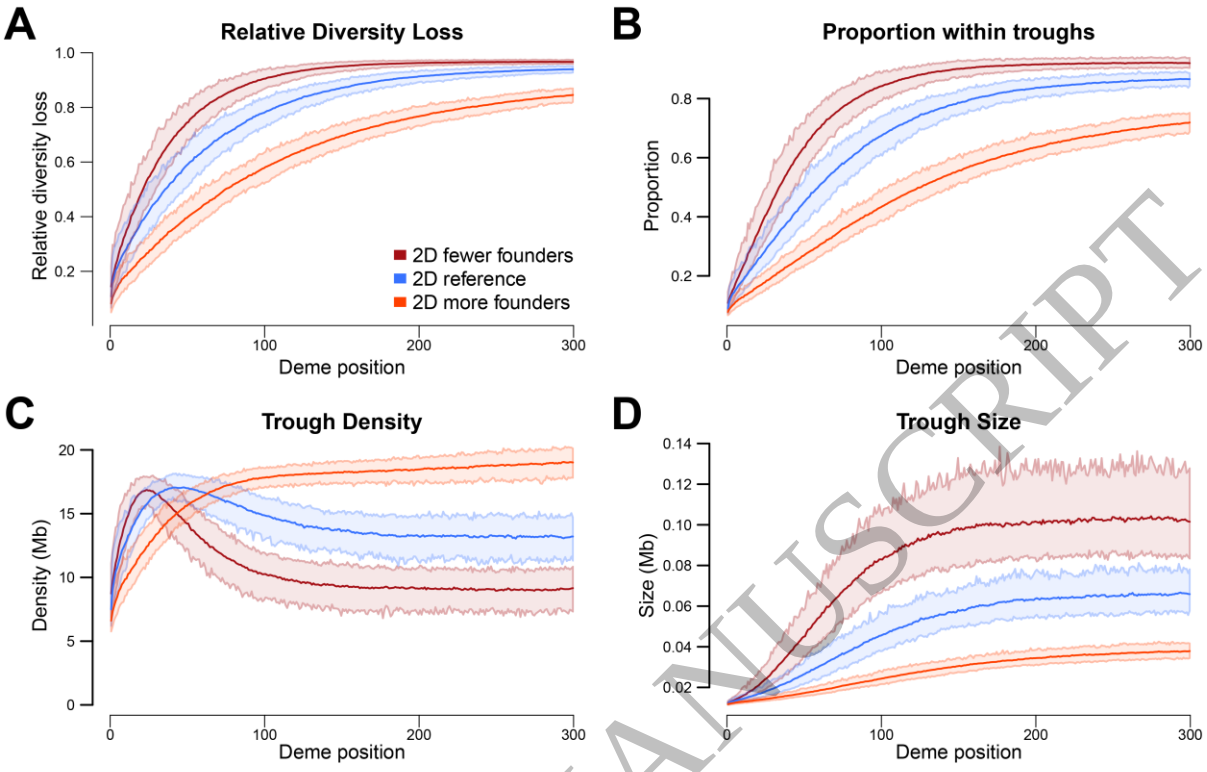

Figure 3
159x100 mm (.41 x DPI)

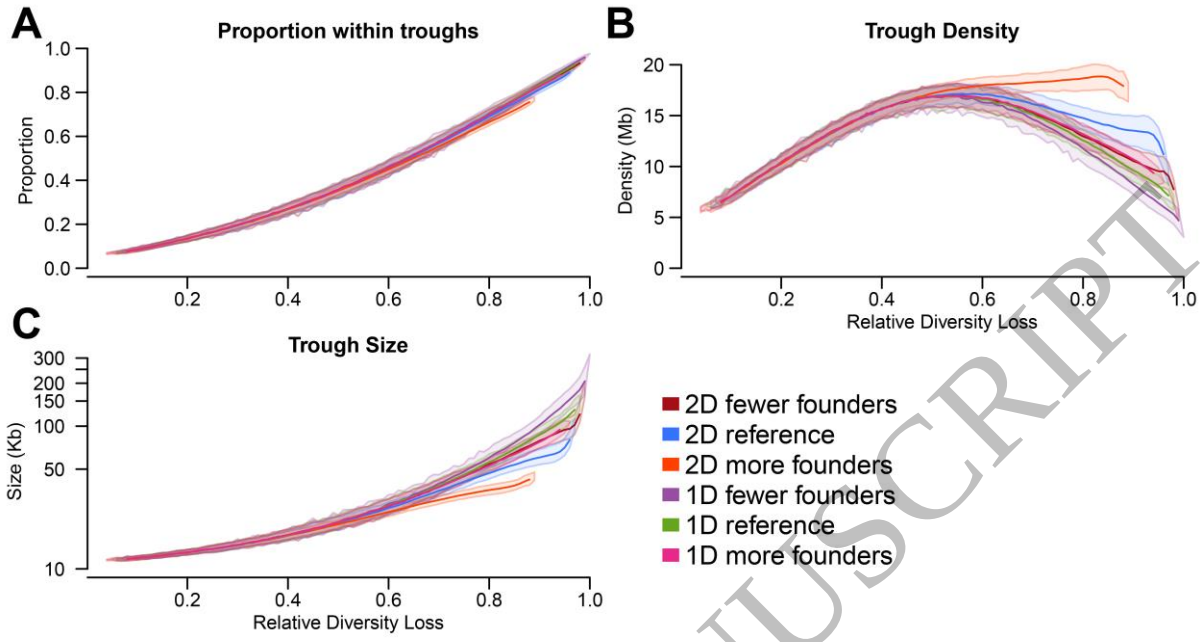

Figure 4
159x85 mm (.41 x DPI)

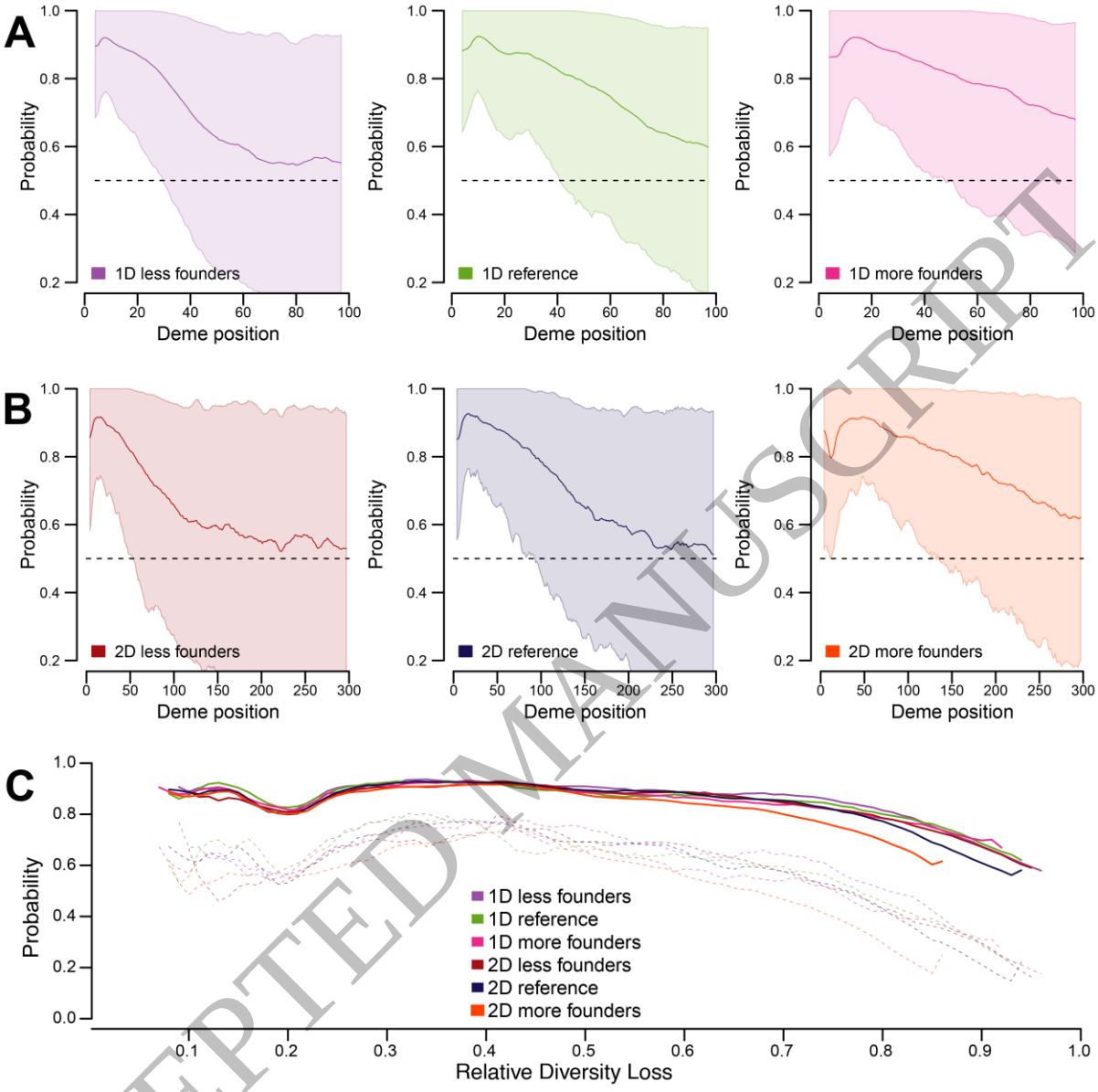

Figure 5
159x158 mm (.41 x DPI)

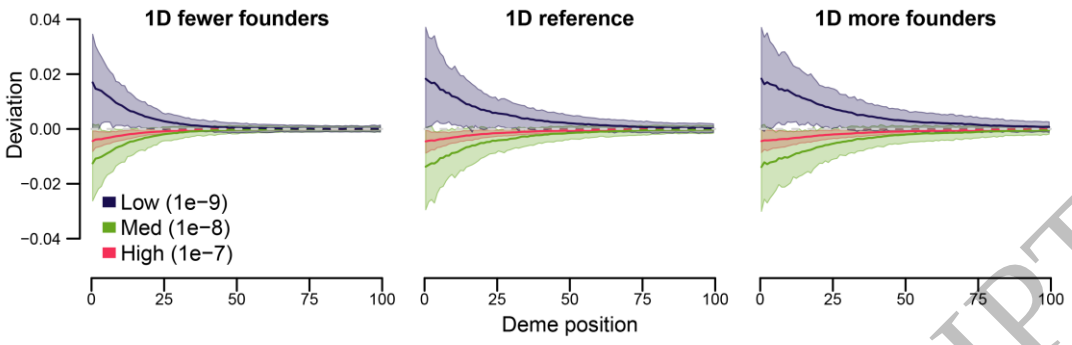

Figure 6
142x44 mm (.41 x DPI)