Title: Natural selection helps explain the small range of genetic variation within species

Authors: Russell B. Corbett-Detig[1], Daniel L. Hartl[1], Timothy B. Sackton[1]*

Affiliations:

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA.

*Correspondence to: tsackton@oeb.harvard.edu

Abstract: The range of genetic diversity observed within natural populations is much more narrow than expected based on models of neutral molecular evolution. Although the increased efficacy of natural selection in larger populations has been invoked to explain this paradox, to date no tests of this hypothesis have been conducted. Here, we present an analysis of whole-genome polymorphism data and genetic maps from 39 species to estimate for each species the reduction in genetic variation attributable to the operation of natural selection on the genome. We find that species with larger population sizes do in fact show greater reductions in genetic variation. This finding provides the first experimental support for the hypothesis that natural selection contributes to the restricted range of within-species genetic diversity.

Main Text:

Understanding the determinants of genetic diversity within populations is a central goal of evolutionary biology, with important implications for the study of such issues as demographic histories of populations (*1*), selective constraints (*2*), and the molecular basis of adaptive evolution (*3*). Under simple neutral models of evolution with constant mutation rates across species, levels of genetic diversity within populations are expected to increase linearly with the number of breeding individuals (the census population size or N_c). However, this prediction has not been borne out in practice. While the range of N_c spans many orders of magnitude, levels of genetic diversity within species fall in a comparatively narrow range (*4-6*). This discrepancy is among the longest standing paradoxes of molecular population genetics (*4,5,7*).

Because under neutral models diversity within a population is determined by the product of the mutation rate and the effective population size, simple explanations for this discrepancy include an inverse correlation between mutation rate and population size (*6,8*) and a greater impact of non-equilibrium demographic perturbations in large populations, such as high variance in reproductive success (*9*) or population-size fluctuations (*10*). However, neither explanation seems sufficient to fully account for the observed patterns of neutral diversity across species (*6*).

Another potential contributor is the operation of natural selection on the genome (5,7,11). Both the adaptive fixation of beneficial mutations (*7*,*12*) and selection against deleterious mutations (*13,14*) purge linked neutral variants via genetic hitchhiking. Furthermore, theoretical arguments (*5,11*,*15*) suggest that, when the impact of natural selection is substantial, the dependence of neutral diversity on population size is weak or non-existent. Therefore, the action of natural selection could provide a resolution to paradox of levels of neutral diversity—if natural selection impacts linked neutral diversity more strongly in populations with a large N_c than in populations with a small N_c. Although many authors have demonstrated the theoretical viability of natural selection as a potential explanation (*6,7,11-13,15)*, no direct empirical tests of this hypothesis have been conducted.

In regions of relatively low recombination, selected variants affect more neutral sites through linkage, and the resulting correlation between recombination and polymorphism (*16*,*17* reviewed in *18*) provides an appealing means of quantitatively assessing the effect of selection on levels of neutral genetic variation within populations (*e.g. 17*) One unique prediction of the natural selection hypothesis is that larger populations will display stronger correlations between recombination and polymorphism than smaller populations. More generally, natural selection will play a greater role in shaping the distribution of neutral genetic variation in species with large N_c. Whole-genome polymorphism data are now available for a wide variety of species (*e.g. 19,20*), and these data enable us to conduct a quantitative, robust test of this prediction of the natural selection hypothesis.

We identified 39 species (16 plants, 6 insects, 2 nematodes, 2 birds, 4 fishes, and 9 mammals) for which a high quality reference genome, a high density, pedigree-based linkage map, and genome-wide resequencing data from at least two unrelated chromosomes within a population were available [Supplemental Methods Section 1, Table S1]. Undoubtedly, our sampling is biased towards more commonly studied clades (e.g., mammals), but this is unavoidable in this type of analysis, and there is no reason in principle why this taxonomic bias would affect the basic conclusions we describe.

After acquiring sequence data, we developed and implemented a bioinformatic pipeline to align, curate, and call genotype data for each species [Figure S1, Supplemental Methods Section 2]. We further used the available genetic maps to estimate recombination rates across the genomes, and then computed partial correlations between recombination rate and neutral polymorphism estimated in 500-kb windows across the genome (excluding sex chromosomes), accounting for differences in gene density [Figure S2, Figure S3, Supplemental Methods Section 3]. Across all species, we analyzed recombination between more than 380,000 markers and aligned more than 57 billion short reads.

Collectively, our data support the hypothesis that natural selection on linked sites eliminates disproportionately more neutral polymorphism in species with large N_c, and in this way natural selection truncates the distribution of neutral genetic diversity. At a coarse scale, there is a stronger correlation between polymorphism and recombination in invertebrates (mean partial tau after correcting for gene density = 0.243), which likely have a large N_c on average, than observed in vertebrates (mean partial tau = 0.131), which likely have smaller N_c on average (two-tailed permutation P = 0.047, Figure 1). We observe similar patterns for herbaceous plants versus woody plants (two-tailed permutation P = 0.021, Figure 1), as well as for alternate window sizes (Supplemental Table S3).

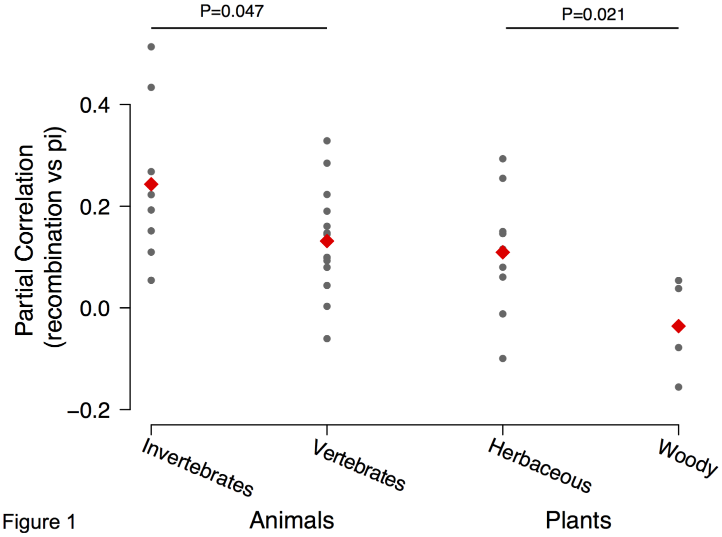

Figure 1. Association between pi and recombination rate is stronger in taxonomic classes expected to have larger population sizes. We estimated both pi and recombination rate in 500-kb windows across the genome and computed partial correlations (Kendall's tau) after accounting for variation in functional density (measured as proportion of sites in a window that are part of an annotated CDS feature). The significance of differences in mean tau (red diamonds) between vertebrates and invertebrates, or between woody and herbaceous plants, was estimated by permutation (100,000 samples) in the statistical programming language R.

Considerable theoretical work in population genetics has been devoted to modeling the impact of natural selection on linked neutral genetic variation (*12,14,21-23*). Building upon this literature, we fit the data with an explicit model relating polymorphism, recombination rate, and density of functional elements in the genome. Using this model, we estimate the fraction of neutral diversity removed by linked selection for beneficial alleles and/or against deleterious alleles [Figure 2, Supplemental Methods Section 4] for each species, as a measure of the degree to which natural selection shapes patterns of neutral diversity within each species. While quantitative estimates of N_c are preferable for testing this hypothesis, in practice it is not feasible to determine N_c for the majority of species we studied. Instead, we used the species' range and body mass as proxies for N_c (Table S2, Supplemental Methods Section 1). We expect that range will be positively correlated with N_c, body mass will be negatively correlated with N_c, and N_c will be positively correlated with the impact of selection.

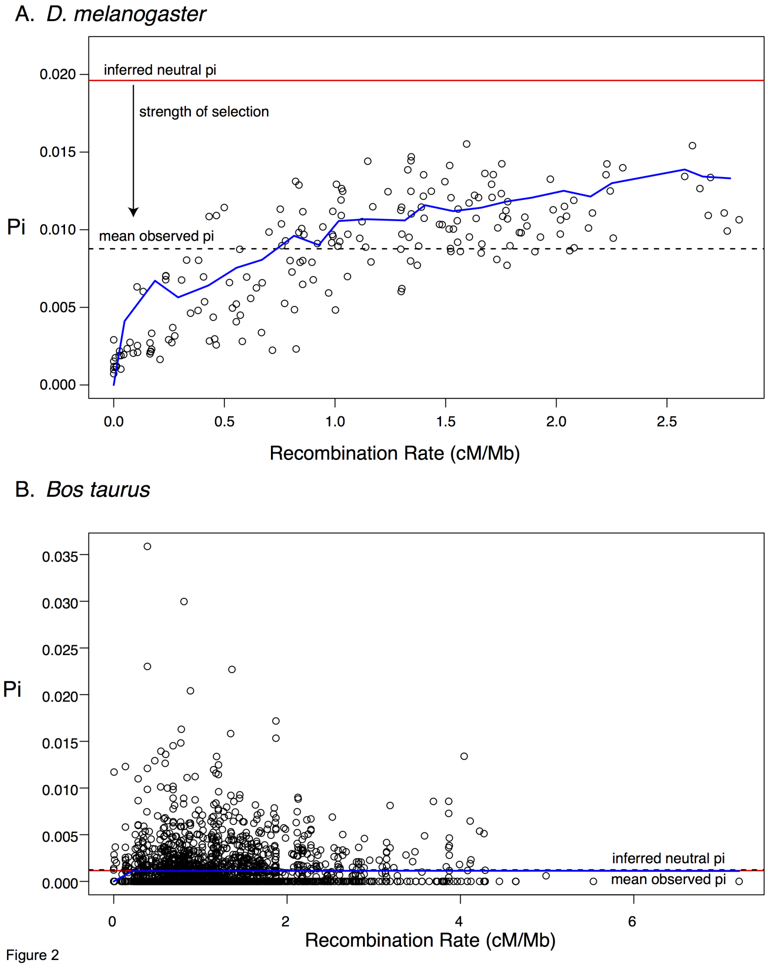

Figure 2. Estimating the impact of selection on linked neutral variation. To obtain a direct estimate of the amount of linked neutral variation removed by selection, we fit a population genetic model incorporating hitchhiking and background selection effects to the estimates of pi and recombination rate in 500 kb windows across the genome. Model fit (blue), estimated neutral pi (red), and observed mean pi (dashed) are shown for a species with large population size (D. melanogaster, part A) and small population size (Bos taurus, part B). The strength of selection is estimated as 1 – (observed pi / neutral pi).

Indeed this is what we observe. For example, in *Caenorhabditis briggsae*, which is the smallest animal that we studied and likely has a very large N_c, we estimate that selection removes 62.4% of neutral genetic variation. In *Bos taurus*, the largest animal included in this study and consequentially likely to have one of the smallest N_c, there is almost no detectable effect of selection patterns of nucleotide variation in the genome.

More generally, we tested the hypothesis that N_c is positively correlated with the impact of selection by fitting a linear model that includes as predictors both body mass and geographic range, and normalized strength of selection as the response variable (Figure 3, Supplemental Methods Section 5). Both mass and range are significant predictors of strength of selection in the expected directions ($\log_{10}$(mass): coefficient = -0.033904, P = 0.000132; log10(range): coefficient = 0.091723, P = 0.028000), and together they explain 43.92% of the variation in strength of selection (overall P = 2.715e-5). Thus, it appears that selection does play a greater role in shaping patterns of neutral genetic variation in species with larger population sizes, and thereby contributes to explaining the narrow range of polymorphism levels observed among species. We obtained similar results when we varied the window size, the details of the selection model, and the approach for normalizing the strength of selection, suggesting that our results are robust to analysis choices [Table S4].

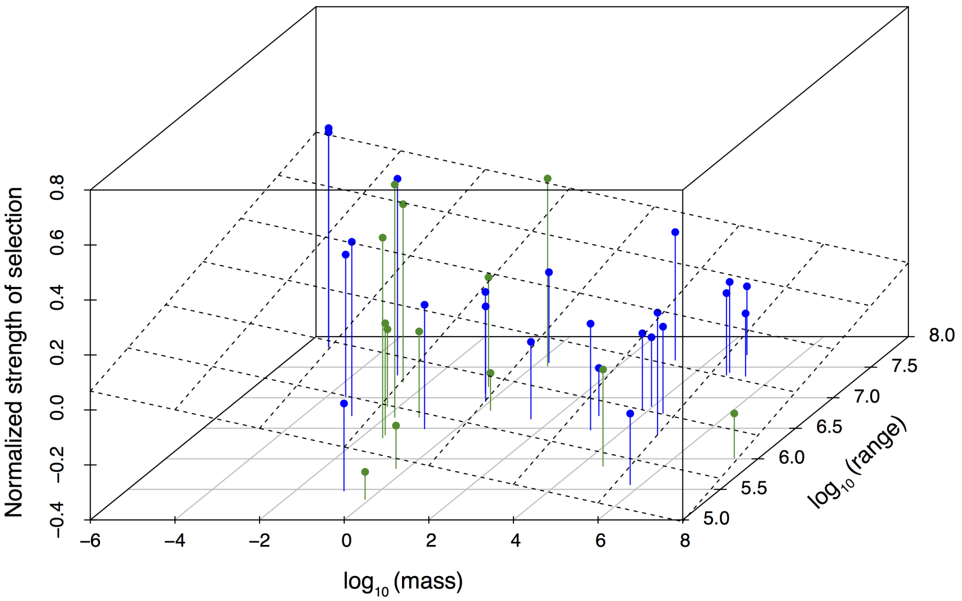

Figure 3

Figure 3. Proxies for census size are correlated with the estimated strength of selection. For each species, we obtained estimates of mass and geographic range, and used those as predictors in a linear model with a normalized measure of the strength of selection as the response. Each point (blue=animals, green=plants) represents a species, and the plane is the regression plane. Species with low mass and large ranges are more impacted by selection.

If the process of recombination is itself mutagenic, neutral processes could produce a correlation between recombination and polymorphism (*16-18*). However, no or very weak correlations between divergence and recombination has been found in most species that have been closely studied [*e.g. 16, 24*, reviewed by *18*]. Moreover, for those species in which a positive correlation between divergence and polymorphism has been found, it is likely the result of selection acting on the ancestral population (*25*). Furthermore, two of the species that showed the strongest correlation between polymorphism and recombination (partial tau = 0.514 for *D. melanogaster*, partial tau = 0.434 for *D. pseudoobscura*) have no such correlation between recombination rate and divergence either on broad scales (*16*) or fine scales (*24*). Therefore differential mutation rates do not appear to be a plausible explanation of the observed patterns.

On the strength of early allozyme polymorphism data, Lewontin (*4*) observed that the range of neutral genetic variation is substantially smaller than the range of N_c. Because both positive and negative selection purge linked neutral mutations, the operation of natural selection on the genome affects patterns of neutral genetic variation at linked sites across the genome. Although many authors have suggested that natural selection may play a role in truncating the distribution of genetic variation (*5-7,11*), no unique test of this hypothesis has been proposed or conducted. Here, we showed that species with larger N_c display a stronger correlation between neutral polymorphism and genetic recombination, and that natural selection removes disproportionately more variation from larger populations. This indicates that natural selection plays a greater role in shaping patterns of polymorphism in the genome of species with large N_c, and these observations confirm that natural selection helps to explain the truncated distribution of neutral genetic variation in natural populations.

Understanding the proximate and ultimate factors that affect the distribution of genetic variation in the genome is a central and long-standing goal of population genetics and it carries important implications for a number of evolutionary processes. One implication of this work is that in species with large N_c, such as *Drosophila melanogaster*, selection plays a dominant role in shaping the distribution of molecular variation in the genome. Among other things, this can affect the interpretation of demographic inferences because it indicates that even putatively neutral variants are affected by natural selection at linked sites. Furthermore, to whatever degree standing functional variation is also affected by selection on linked sites (*e.g. 22*), local recombination rate in organisms with large N_c may also predict what regions of the genome will contribute the greatest adaptive responses when a population is subjected to novel selective pressures. In its broader implications, this work provides some of the first direct empirical evidence that the standard neutral theory may be violated across a wide range of species.

Methods

1. Data sources and curation

Reference genome versions, annotation versions, map references, and other basic information about the genetic and genomic data for species we included in our analysis is summarized in Supplemental Table S1, and described in more detail below.

Reference genomes

To identify suitable species for our analysis, we started from the list of genome projects available at GOLD (http://www.genomesonline.org/documents/Export/gold.xls) and NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/), both accessed 6 October 2013. We removed all non-eukaryotes from both sets. We then further filtered the GOLD set to remove all projects where status was not either “draft” or “complete”, and where project type was not “Whole Genome Sequencing”, and the NCBI set to keep only all projects with MB > 0 and status equal to “scaffold,” “contigs,” or “chromosomes.” Finally, we merged both lists, removed duplicate species, and removed all species without an obligate sexual lifestyle.

Next, we manually checked the quality of the genome assembly of each species remaining on our list by inspection of assembly reports available from NCBI, Ensembl, Phytozome, or species-specific databases. Any species without chromosome-scale assemblies was removed, as was any species without an available annotation of coding sequence. In two cases (*Heliconius melpomene* and *Gasterosteus aculeatus*), chromosome scale assemblies are available but annotations were only available for the scaffold-level (or a previous, lower quality chromosome-level) assembly. In these cases, we updated the coordinates of the coding sequence annotations using custom Perl scripts, available upon request from the authors.

Polymorphism

We required that each species be represented by random-shearing Illumina short-read sequence data for at least two chromosomes derived from unrelated individuals from within the same population. In some cases, this means that a single outbred diploid individual was used. If samples were intentionally inbred or if the species is known to engage in frequent self-fertilization in natural populations, we required data from at least two separate individuals. In some cases, we used polymorphism data from a very closely related taxon to the genome species. In particular, we attempted to avoid using polymorphism data from domesticated species where possible; in many cases we were able to use polymorphism data from wild ancestors or close relatives of domesticated plants and animals.

Genetic maps

We required that each species have available a pedigree-based genetic map, generated from markers that could be mapped to the reference genome by either ePCR or BLAST, and with an average inter-marker spacing (after filtering unmapped and mis-mapped markers; see below) of no more than 10 cM. For species with recombination in both sexes, we used sex-averaged genetic distances where possible, although in a few cases maps were only available for a single sex. For species with recombination in only one sex, we corrected genetic distances to represent a sex-averaged value by dividing by 2. As with polymorphism data, we occasionally used genetic maps from a closely related taxa, particularly in cases where substantially higher quality maps were available.

Range and mass information

While ideally we would obtain estimates of actual census population sizes, even moderately accurate estimates are rarely available. As an alternative, we used species range and species mass as proxies for census population size.

To determine range, we estimated species distribution from the sources listed in Supplemental Table 2 and computed the square kilometers of the distribution using a combination of Google Maps API to estimate the areas of polygons drawn on maps (as implemented here: http://www.freemaptools.com/area-calculator.htm) and Wolfram Alpha to compute the area of defined regions. Further details are supplied in Supplemental Table 2.

For most animals, estimates of mass are available from published references (Supplemental Table 2). However, mass data for plant species is generally not readily available. Thus, to estimate a comparable value for plants, we obtained estimates of average plant height from published sources (Supplemental Table 2). We converted height to mass by assuming each plant can be approximated as a cylinder with a density estimated based *A. thaliana* for herbaceous species, and assumed to be 0.5 g/cm^3 for woody species. Both height and mass values are reported for plants (Supplemental Table 2).

2. Polymorphism pipeline

Alignment and genotyping pipeline

Our alignment and genotyping pipeline is summarized in Supplemental Figure S1, and described in detail below.

We acquired short read data from the NCBI short read trace archive. All accession numbers for short read data used in this analysis are listed in Supplemental Table S5. We aligned these data to their respective reference genomes (reference genome versions are listed in Supplemental Table S1). For libraries prepared from genomic DNA we used bwa v0.7.4 (*26*). For libraries prepared from RNA, we aligned reads initially using tophat2 v2.0.7 (*27*). For both DNA and RNA we then realigned reads that failed to align confidently using Stampy v1.0.21 (*28*). After this, putative PCR duplicates were removed from both RNA and DNA based libraries using the 'MarkDuplicates' function in Picard v1.98 (http://picard.sourceforge.net). For DNA libraries, we next use the ‘indelRealigner’ tool in the GATK v2.4-3 (*29*) to realign reads surrounding likely indel polymorphisms. All programs were run using default parameters, except that we used the ‘--no-novel-juncs’ and ‘--no-coverage-search’ options in tophat2.

We genotyped all samples using the GATK v2.4-3 (*29*). If samples were intentionally inbred, or if the species is known to primarily reproduce through self-fertilization in natural populations we used the ‘-ploidy’ option to set the expected number of chromosomes to 1 (see Supplemental Table S1 for ploidy settings used for each species). We then extracted polymorphism data from four-fold degenerate synonymous sites. While there is mounting evidence that these sites are not evolving under strictly neutral processes (*e.g. 30, 31*), four-fold degenerate sites are a widely accepted approximation for neutral markers in the genome, and importantly these sites are available in both RNA and DNA sequencing efforts.

We sought to exclude low confidence sites by filtering our genotype data through several basic criteria. First, we required that every fourfold degenerate site have a minimum quality of 20 computed across the entire sample. Second, for every fourfold degenerate site, we computed the mean depth for each sample. We then required each sample have at least half as many reads as the mean depth at a site for that position to be included in the analysis. For variable sites, we further required that phred-scaled strand bias be below 40, and the absolute value for the Z-score associated with the read position rank sum, the mapping quality rank sum, and the base quality rank sum be above 4. See the GATK (*29*) documentation for in-depth descriptions of the relevant filters used. We applied these criteria to both DNA and RNA based libraries. Summaries of sites aligned and filtered for each genome are available in Supplemental Table 6.

Homo sapiens

Rather than recompute variant calls, for the human data, we obtained VCF files for the Yoruban population from (*20*). We elected to do this because these data are exceedingly well curated and the size of the human variation raw data presents a practical computational challenge. The VCF file was treated as described below in all case.

Estimating pi in genomic windows

From these filtered files, we computed average pi in non-overlapping windows of 100kb, 500kb, or 1000kb. In all cases, we excluded windows from our analysis with fewer than 500 sequenced four-fold degenerate sites. We also exclude all windows on sex chromosomes, in order to avoid complicating effects of hemizygosity on patterns of polymorphism.

3. Recombination rate estimation pipeline

We did not attempt to recompute genetic maps from raw genotype data, as in most cases this raw data is unavailable. Therefore, our standard approach to estimating recombination rates is to first obtain sequence information and genetic map positions for markers from the literature, map markers to the genome sequence where necessary, filter duplicate and incongruent markers, and finally estimate recombination rates from the relationship between physical position and genetic position. This pipeline is summarized in Supplemental Figure 2, and described in more detail below. Specific details for each species are described in supplemental text S1.

Data curation and mapping markers to the reference genome

We used three basic approaches to link markers from genetic maps to sequence coordinates. In some cases, sequence coordinates are available from the literature, in which case we use previously published values (in some cases updated to the latest version). For cases where primer information (but not full sequence information) is available, we used ePCR (*32*) with options -g1 -n2 -d50-500 and keeping all successful mappings, except where noted. For cases where locus sequence information is available, we used blastn with an e-value cutoff of 1 x 10^{-8} and retain the top 8 hits for each marker, except where noted. In both cases, we only retain positions where the sequence chromosome and the genetic map chromosome are identical. Specific curation and data cleaning steps for individual species are summarized in Supplemental Table S7.

Removal of incorrectly ordered or duplicated markers

For most species, the genetic position and physical position of markers along a chromosome are not completely congruently ordered. That is, physical position is typically not strictly monotonically increasingly with genetic position. Incongruent markers can arise from incorrect genome assemblies, errors in map construction, or sequence rearrangements between the reference genome and the mapping population.

For consistency, we assume that the reference genome is correctly assembled, and we correct the order and orientation of genetic maps to be consistent with the sequence assembly. To remove incongruent markers, we find the longest common subsequence (LCS) of ranked genetic and physical positions, and define as incongruent all markers that are not part of the LCS. After removing incongruent markers, we filtered each map to retain only the single most congruent mapping position for markers with multiple possible genomic locations. Functions to perform this analysis in R are available from T. B. S. upon request.

Masking low quality map regions

To improve the quality of our recombination rate estimation, we designed a masking filter to exclude regions of chromosomes where the fit between the genetic map and the physical position of markers is particularly poor, defined as a run of 5 bad markers (for chromosomes with at least 25 markers), or a run of 0.2 times the number of markers on the chromosome, rounded up, bad markers (for chromosomes with at fewer than 25 markers). We also completely mask any chromosome with fewer than 5 markers in total.

The final map quality and various filtering results are summarized in Supplemental Table S8.

Recombination rate estimation

Our basic approach to recombination rate estimation is to fit a continuous function to the Marey map relating genetic position and physical position for each chromosome. We use two different approaches that result in different degrees of smoothing: a polynomial fit and a linear B-spline fit. In both cases, we start by optimizing the polynomial degree or spline degrees of freedom using a custom R function that maximizes the Akaike Information Criteria (AIC) for the model fit. For the polynomial fit, we optimize between degree 1 and degree max(3, min(20, # markers / 3)). For the B-spline fit, we optimize degrees of freedom between df 1 and min(100, max(2, #markers/2)). In each case, we retain the value with the highest AIC. To compute recombination rates in cM/Mb, we then take the derivative of the fitted function, evaluated at the midpoint of each window. For additional smoothing, we set all values of recombination estimated below 0 to 0, and all values above the 97.5th percentile to the 97.5th percentile. While the two estimates tend to be highly correlated with each other, the polynomial fit appears to perform better for low quality maps, and the B-spline fit for high quality maps. Therefore, unless otherwise noted, we use the polynomial estimates of recombination rate for maps with inter-marker spacing of greater than 2 cM, and the B-spline estimates for maps with inter-marker spacing less than or equal to 2 cM. All estimation was done in R; code is available upon request.

An example of the map fitting procedure is shown in Supplemental Figure S3.

Partial correlations between recombination rate and pi

To estimate the strength of the association between recombination rate and pi, we use partial correlations that account for variation in coding sequence density across the genome. In many species (*22, 33, 34*) recombination rate and/or neutral diversity is correlated with gene density, and thus we need to account for this confounding variable in our analysis. We do this using partial correlations, implemented with the ppcor package in R.

First, we estimate coding sequence density in each window as the fraction of each window represented by CDS sites, extracted from the same GFF files for each species used to compute four-fold degenerate sites. We then estimate Kendall's tau between recombination rate and pi for each window after correcting for coding sequence density.

4. Modeling the joint effects of background selection and hitchhiking on neutral diversity

Ultimately, our goal is to infer the strength of the impact of natural selection on linked neutral diversity across species. To go beyond correlation analysis, we build upon the extensive literature in theoretical population genetics. Two primary types of selection can introduce a correlation between recombination rate and pi: background selection (BGS) and hitchhiking (HH). Here, we are not concerned with distinguishing between the two models, and so focus on their joint effects.

We begin with the very general selective sweep model derived by Coop and Ralph (*21*), which captures a broad variety of hitchhiking dynamics. To include the effects of background selection, we rely on the fact that to a first approximation, background selection can be thought of as reducing the effective population size and therefore increasing the rate of coalescence. This effect can be incorporated into the (*21*) model by a relatively simple modification to equation 16 of Coop and Ralph *(21).* Specifically, we scale *N* by a BGS parameter, exp(-*G*), in equation 16, which then leads to a new expectation of pi:

E[pi] = theta / [1/exp(-*G*) + alpha / rbp], [Equation 1] where alpha = $2N$ * Vbp * J2,2 (per *21*) and rbp is the recombination rate per base pair.

This is very similar to previously published models of the joint effects of background selection and hitchhiking (*e.g. 23)*

In addition to combining background selection and hitchhiking, we would also like to relax the assumption that these processes act uniformly across the genome. All else being equal, regions of the genome with a higher density of potential targets of selection should experience a greater reduction in neutral diversity. To account for this variation in the density of targets of selection, we build upon the approach of Rockman *et al.* (*22*), which derives from the work of (*14*). Specifically, we fit the following model to estimate G for each window k:

$$Gk = \mathrm{sum}[U{*}fd(i){*}sh \,/\, 2(sh{+}P|Mk - Mi|)(sh + P|Mk - Mi + 1|)], \qquad \text{[Equation 2]}$$

where U is the total genomic deterious mutation rate, $fd(i)$ is the functional density of window i, sh is a compound parameter capturing both dominance and the strength of selection against deleterious mutations, Mk and Mi are the genetic positions in Morgans of window k and window i respectively, and P is the index of panmixis, which allows us to account for the effects of selfing (*22*, *34*). We estimate functional density as the fraction of exonic sites in the genome that fall within the window in question. We focus on exonic sites as a proxy for targets of selection as they are the only functional measure that is uniformly available for all the species in our study.

Because *P*, *U*, and *sh* are not known, we fit this BGS model with a variety of parameter combinations. For *U*, we fit both a model where we estimate *U* [Ubest] as the mutation rate times the number of exonic bases in the genome, and a model where we assume a maximum U of max(1, 2*Ubest). For *P*, we assume 1 for all vertebrates, insects, and obligate outcrossers among plants; 0.04 for highly selfing species, and 0.68 for partial selfers. These estimates correspond to selfing rates of 0%, ~98%, and ~50%, respectively. Estimates of selfing are available in Supplemental Table S9. For a few species of plants we were unable to obtain reliable estimates of selfing rate (indicated by NA in Supplemental Table S9), and thus we do not include these species in our modeling approach. For *sh*, we fit a range of values evenly spaced (on a log scale) between 1e-5 and 0.1. Code to estimate *Gk* was implemented in C++ and is available upon request from the authors.

To incorporate functional density into the hitchhiking component of the model, we make the simplifying assumption that sweeps targeting selected sites outside a window will have little effect on neutral diversity within a window, and that sweeps occur uniformly within a window. Under this assumption, we can consider functional density as a scaling factor on the rate of sweeps, Vbp. Specifically, we reparameterize the rate of sweeps, Vbp, as V, the total sweeps per genome, and then consider the fraction of sweeps that occur in a particular window i as V*fd(i). This results in a simple scaling of alpha in equation 1.

Incorporating the effects of functional density in both BGS and HH, our final model for the expectation of neutral diversity in window [i] is:

$$E(pi[i]) = theta[neutral] / (1/exp(–G[i]) + alpha*fd[i] / rbp[i]), \qquad \text{[Equation 3]}$$

To obtain an estimate of the effect of selection for each species, we fit this model for estimates of G[i] derived from different parameter combinations (see above), using the nlsLM() function from the minpack.lm package in R. In addition, we fit a background selection only model, in which alpha=0 and thus the second half of the denominator is 0.

From each model fit we estimate theta[neutral] for both the joint model and the background selection only model, and also extract the likelihood of the fit. We then compute the AIC for each parameter combination, and use the AIC to estimate the relative likelihood of each model *j* as

$$RELj = exp((AICmin - AICj)/2) \qquad \text{[Equation 4]}$$

which we then normalize so that the relative likelihoods for all models for a species sum to 1.

We then estimate neutral pi for each species as the weighted mean of the theta[neutral] estimates, weighted by their relative likelihoods. We then compute average observed pi for each species, and report as the strength of selection 1 – (avgpi.observed/pi.neutral). Values below 0 are replaced by 0. This value can be interpreted as the proportion of neutral variation removed by selection acting on linked sites, averaged across the genome.

5. Linear models

Our goal is to test whether N_c predicts the degree to which selection shapes patterns of neutral diversity, using log-transformed measures of body mass and geographic range as proxies for N_c. However, many other factors could potentially influence our measure of strength of selection, including biological factors such as genome size, average recombination rate, and taxonomic class; and experimental factors such as map quality, assembly quality, and polymorphism quality. In particular, we might expect to underestimate the strength of selection in species with low-quality assemblies, maps, or polymorphism datasets, and we might expect that on average larger genomes and higher recombination rates would reduce the impact of selection.

In order to account for these parameters that are not directly of interest, and at the same time avoid overfitting the model, we used the R package glmulti to exhaustively search all possible main effect combinations of predictors from this set: log(mass), log(range), kingdom (animal or plant), genome size (estimated C-value), average recombination rate, fraction of total assembly that is gaps, scaffold N50 normalized by average chromosome size, average intermarker spacing of good markers on the genetic map, fraction of total markers that are retained as useable, and the proportion of 4D sites filtered from the resequencing data. We used AICc as the information criteria, and selected the best-fitting model overall to use (but we obtain identical results if we select the terms with weighted importance > 0.5 based on the top 10 models).

We obtain assembly information from NCBI, Phytozome, or the original genome publication. C-values for plants come from http://data.kew.org/cvalues/, and C-values for animals come from http://www.genomesize.com/. In all cases, most recent estimates, "prime" estimates, or flow cytometry estimates are preferred; where several seemingly equally good estimates are available, the average is used. In some rare cases a related species is used instead of the sequenced species if the sequenced C-value is not available.

Normalized N50 is computed as scaffold N50 / average chromosome size, including sex chromosomes (and separate arms for Dmel, Dpse, and Agam). In a few cases where scaffold N50 is unavailable but chromosomes are assembled into single scaffolds, median chromosome size is used as scaffold N50. For polymorphism quality, Hsap is from 1000 genomes, so we assume very high quality (.99). Average recombination rate computed as the mean of the piece rate or poly rate for the 500kb window data, except for Sbic which is the median because of outliers in the polynomial fit. All raw data is available as Supplemental Table S10.

The best model is

selection strength ~ log(mass) + log(range) + kingdom + fraction of total markers that are retained as useable

For model fitting, we want to test the effects of log(mass) and log(range) after correcting for kingdom and fraction of total markers that are retained as useable; to do this, we first fit a model with only the latter two predictors, and retain the residuals, and then use log(mass) and log(range) as predictors against the retained residuals. Thus, our final model is:

residuals(lm(selstr ~ kingdom + prop.good) ~ log(mass) + log(range).

To assess robustness, we use several alternate approaches, including different window sizes, selecting the best model by AIC (instead of a weighted average of neutral pi), and using either no normalization or all quality parameters (instead of just the best fit model). In all cases our main conclusions are robust to these analysis choices (Supplemental Table S4).

References and Notes:

1. J.E. Pool, I. Hellman, J.D. Jensen, R. Nielsen. Population genetic inference from genomic sequence variation. *Genome research* **20.3**:291-300 (2010).

2. M. Kellis et al. Defining functional DNA elements in the human genome. *PNAS* 201318948 (2014).

3. R. Nielsen. Molecular signatures of natural selection. *Annu. Rev. Genet.* 39:197-218 (2005).

4. R.C. Lewontin, The genetic basis of evolutionary change. (Columbia University Press, New York, xiii).

5. J.H. Gillespie. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155.2: 909-919 (2000)

6. E.M. Leffler et al. Revisiting an old riddle: what determines genetic diversity levels within species?. *PLoS biology* 10.9: e1001388 (2012)

7. J. Maynard-Smith, J. Haigh. The hitch-hiking effect of a favourable gene. *Genetical research* 23.01: 23-35 (1974)

8. M. Lynch. The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics* 180.2: 933-943 (2008).

9. P. Hedrick. Large variance in reproductive success and the Ne/N ratio. *Evolution* 59.7: 1596-1599 (2005).

10. J.A. Vucetich, T.A. Waite, L. Nunney. Fluctuating population size and the ratio of effective to census population size. *Evolution*. 2017-2021 (1997).

11. J.H. Gillespie. Is the population size of a species relevant to its evolution?. *Evolution* 55.11: 2161-2169 (2001).

12. N.L. Kaplan, R. R. Hudson, C. H. Langley. The "hitchhiking effect" revisited. *Genetics* 123.4: 887-899 (1989).

13. B. Charlesworth, M.T. Morgan, D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134.4: 1289-1303 (1993).

14. R.R. Hudson, N. L. Kaplan. The coalescent process and background selection. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 349.1327: 19-23 (1995).

15. R.A. Neher, T.A. Kessinger, and B.I. Shraiman. Coalescence and genetic diversity in sexual populations under selection. *Proceedings of the National Academy of Sciences* 110.39: 15836-15841 (2013).

16. D.J. Begun, and CF. Aquadro. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* :519-520 (1992).

17. J.J. Cai et al. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS genetics* 5.1: e1000336 (2008).

18. A.D. Cutter, B.A. Payseur. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* 14.4: 262-274 (2013).

19. C.H. Langley, et al. Genomic variation in natural populations of Drosophila melanogaster. *Genetics* 192.2: 533-598 (2012).

20. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491.7422: 56-65 (2012).

21. G. Coop, P. Ralph. Patterns of neutral diversity under general models of selective sweeps. *Genetics* 192.1: 205-224 (2012).

22. M.V. Rockman, S. S. Skrovanek, and L. Kruglyak. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* 330.6002: 372-376 (2010).

23. Y. Kim, W. Stephan. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155.3: 1415-1427 (2000).

24. S.E. McGaugh, et al. Recombination modulates how selection affects linked sites in Drosophila. *PLoS biology* 10.11: e1001422 (2012).

25. D.J. Begun et al. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS biology* 5.11: e310 (2007).

26. H. Li, R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 25:1754-60 (2009).

27. D. Kim *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14(4):R36 (2013).

28. G. Lunter, M. Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*. 21(6):936-9 (2011).

29. M. DePristo *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 43:491-498 (2011).

30. D.C. Shields *et al*. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular biology and evolution* 5.6: 704-716 (1998).

31. D.S. Lawrie *et al*. Strong purifying selection at synonymous sites in D. melanogaster. *PLoS genetics* 9.5: e1003527 (2013).

32. G.D. Schuler. Sequence mapping by electronic PCR. *Genome Res.* 7(5):541-50 (1997).

33. M. Nordborg *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS biology* 3.7: e196 (2005).

34. J.M. Flowers *et al.* Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Molecular biology and evolution* 29.2: 675-687 (2012).

Supplemental Tables (included at end of text)

Supplemental Figures

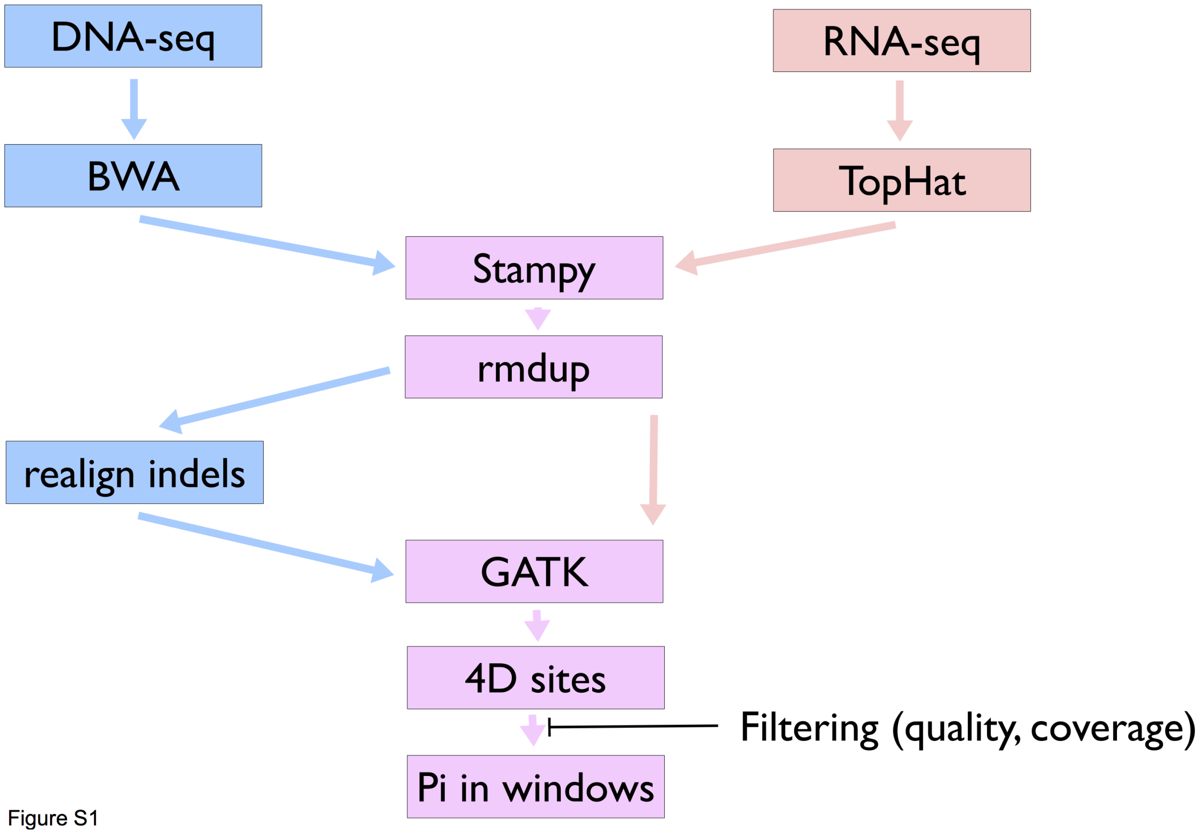

Figure S1

Figure S1. Overview of genotyping pipeline. The pipeline steps to produce final genotype calls and per-site estimates of pi across the genome are summarized, and described in more detail in Supplemental Methods, Section 2.

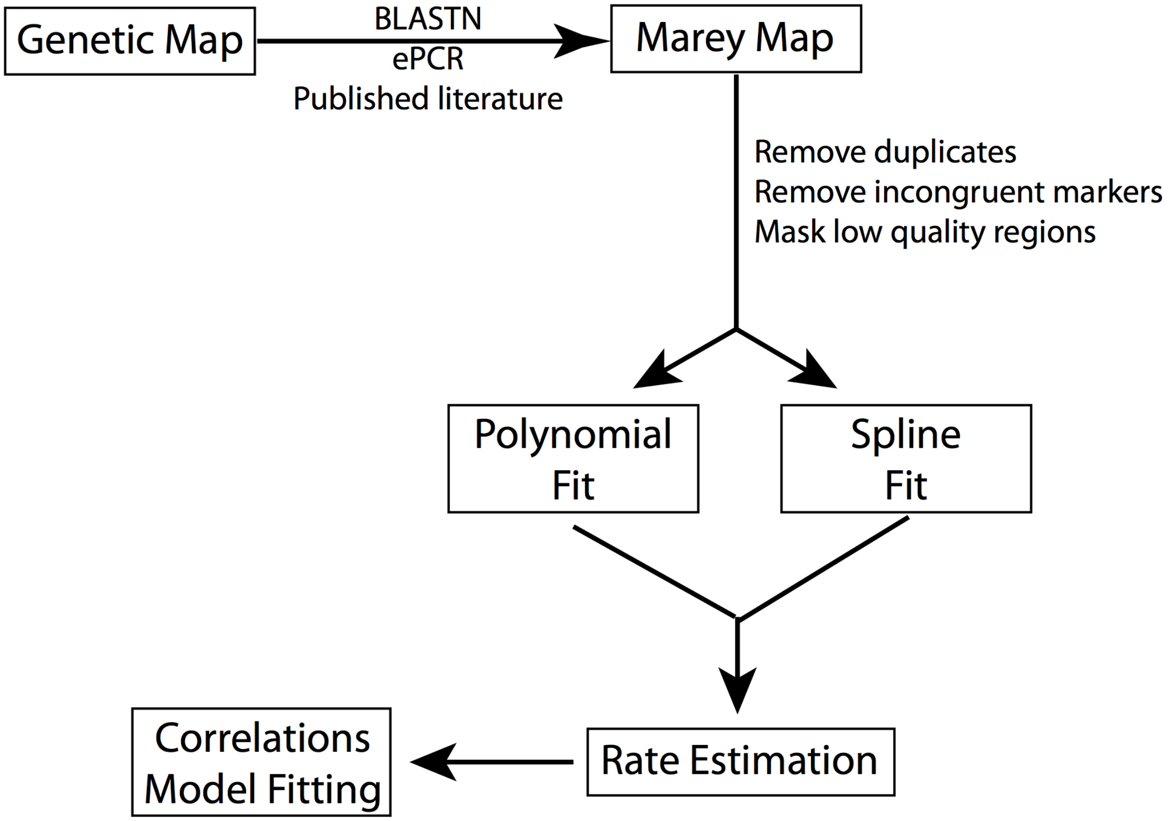

Figure S2

Figure S2. Overview of recombination pipeline. The pipeline steps to produce estimates of recombination rate from initial genetic map data are summarized, and described in more detail in Supplemental Methods, Section 3.

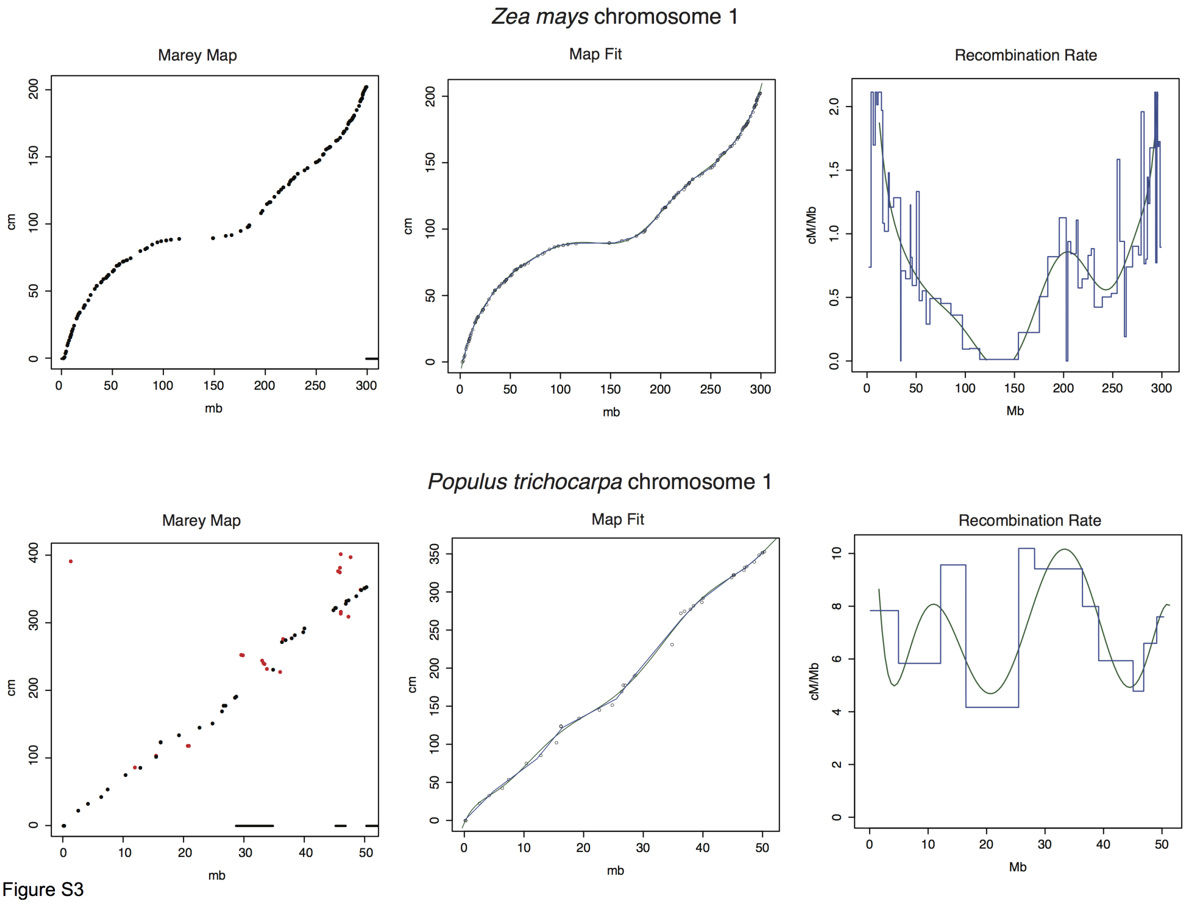

Figure S3

Figure S3: Example of map fitting and recombination estimate pipeline, showing a high quality map (part A) and a moderate quality map (part B). Panels indicate, from left to right: the Marey map relating cM to Mb (with incongruent markers in red and masked regions indicated with solid bars near the x axis); the fit of the polynomial (green) and linear B-spline (blue) functions; the estimate of recombination rate from the polynomial (green) and linear B-spline (blue) functions.

Supplemental Text S1: Genetic map construction details

Anopheles gambiae
Map and markers are derived from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207350/. Mapped to AgamP3 assembly by ePCR using standard approach described below. Chromosome arms were converted to whole chromosomes for consistency with the map by adding the sequence length of the R arms to the L arms based on VectorBase numbers for chromosome length.

Apis mellifera
Map and marker sequences obtained from the map reference, and mapped to the latest Amel assembly via ePCR.

Arabidopsis thaliana
Map and markers derived from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1564425/. Genomic locations updated to latest assembly (TAIR10) by taking the midpoint of the coordinates for the gene id for each marker (from Supplemental Table 5), based on the mapping data available at ftp://ftp.arabidopsis.org/home/tair/Maps/mapviewer_data/TAIR9_AGI_gene.data.

Bombyx mori
Map and markers derived from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2395255/. Mapped to integretedseq assembly from KAIKOBase by ePCR using standard approach described below.

Bos taurus
Map, markers, and positions derived from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2680908/. Converted from Btau4.0 to UMD 3.1 via UCSC LiftOver, with default options. Markers where coordinates were not successfully transferred were removed.

Brachypodim distachyon
Map, markers, and positions derived from supplemental material associated with http://www.ncbi.nlm.nih.gov/pubmed/21597976

C. briggsae
Map and positions come from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3136444/. Currently using the F2 map, not the advanced intercross map, as that avoids having to deal with issues of map expansion.

C. elegans
Data downloaded from WormBase at ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/annotation/genetic_limits/c_elegans.current.genetic_limits.gff2.gz on 11/21/2013. Only cloned, absolute position markers are used.

Canis lupus
Map and markers derived from http://www.genetics.org/content/184/2/595.full. CanFam2 positions published as supplemental material were converted to CanFam 3.1 positions using UCSC LiftOver, with default options. Markers where coordinates were not successfully transferred were removed.

Capsella rubella
Map, marker, and positions are derived from the genome paper via personal communication.

Citrullus lanatus
Map, markers, and positions come from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3256148/. Scaffold positions were converted to chromosome positions using the supplemental table of scaffold positions on chromosomes from the map reference.

Citrus clementina
Map and markers are derived from the *C. clementina* map described in http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3546309/. Markers were mapped to the reference genome via BLAST as follows:
1) Accessions for each marker were retrieved from the supplemental material in the mapping paper
2) FASTA sequence for each accession was retrieved from Genbank using Batch Entrez agains the GSS, Nucleotide, and EST databases in turn.
3) FASTA sequences were concatenated and blasted against the reference genome with an E-value cutoff of 1e-08 and a max output number of 5 sequences.
4) For each marker query sequence, we kept the single best hit and used the start position of the best HSP as the genomic coordinates of the marker.

Cucumis sativus
Map and markers are derived from http://www.ncbi.nlm.nih.gov/pubmed/22487099. Mapped to the reference genome via standard ePCR.

Danio rerio
Map, markers, and positions taken directly from supplemental information associated with http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178105/

Drosophila melanogaster
Map, markers, and positions are derived from Comeron's data (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3469467/), but converting to a cM map (based on assuming c in cM/Mb is constant over each 100kb window, the next midpoint will be c*0.01 cM from the previous midpoint in cM).

Drosophila pseudoobscura

Map, markers, and positions are derived from Noor's data (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3496668/), using the Chromosome 2 map (uncondensed) from the Flagstaff population.

Equus caballus:
The horse map is a sex-averaged map derived from two mapping families described in http://www.ncbi.nlm.nih.gov/pubmed/16314071. Data obtained from ArkDB (http://www.thearkdb.org/arkdbgridqtl/SelectAnalysis.action?speciesid=ARKSPC00000005), and one error was corrected (Marker NVHEQ232 mislabeled as NVHEW232).

To assign markers to chromosomal locations in the genome sequence, the following procedure was used:
1) All markers that could be identified in the UniSTS map of horse available from ftp://ftp.ncbi.nih.gov/genomes/MapView/Equus_caballus/sequence/BUILD.2.2/initial_release/ were assigned to the start position of the location indicated in the MapView file. In all cases below the start position (lowest genomic coordinate) was used for consistency.
2) Of the remaining markers, the location was taken from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2685479/ for any markers assigned a genomic location in that reference.
3) Remaining markers were mapped by ePCR to the horse genome using standard approaches.
4) For a subset of markers where full sequence accessions could be identified, location was determined by blast against the horse genome using the sequence from Genbank.

Primers and other marker information came from:
1) www.ncbi.nlm.nih.gov/pmc/articles/PMC2587302/ (supplemental table and .md file from Genbank, Equus map viewer)
2) other primer sequences identified from:
http://www.uky.edu/Ag/Horsemap/Maps/LINDGREN.PDF
http://dga.jouy.inra.fr/cgi-bin/lgbc/main.pl?BASE=horse
3) references in ttp://www.ncbi.nlm.nih.gov/pubmed/16314071

Gallus gallus
Markers, map, and locations were derived from the supplemental material associated with http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2661806/. Converted from WASHUC2 (build 2.1) coordinates to Ggal4.0 coordinates via UCSC LiftOver, with default options.

Gasterosteus aculeatus
Markers, map, and locations derived from supplemental material associated with http://www.ncbi.nlm.nih.gov/pubmed/23601112.

Glycine max
Markers, map, and locations derived from the Glycine Max Consensus 4.0 map available from Soybase (http://www.soybase.org/dlpages/index.php) and described in http://naldc.nal.usda.gov/catalog/4252.

Gossypium raimondii
Diploid D genome map positions (cM) from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1470701/, available as a spreadsheet here: http://www.plantgenome.uga.edu/cottonmap.htm. Marker mapping to genome sequence based on BLAST (same settings as ccle) using sequences from the map paper, or the cotton genome database (http://www.cottongen.org/data/download) when sequences were not available in map paper.

Heliconius melpomene
Markers, maps, and locations are derived from (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398145/), with the help of personal communications from John Davey. Chromosomal sequences were assembled from the AGP file available at http://www.butterflygenome.org/node/4.

Homo sapiens
Use deCODE genetics (Kong et al 2010) male and female maps, but compute sex-averaged map in R for usage. Using the overall map, not the map broken down by carrier status. Map, markers, and positions available as supplemental data (http://www.decode.com/addendum/). Coordinates updated to hg19 via UCSC LiftOver.

Lepisosteus oculatus
Map and marker sequences are from a personal communication. Marker sequences were mapped to the genome via standard BLASTN.

Macaca mulatta (mmul):
The map was assembled from data available at the website of the Southwest Primate Research Center, here: http://baboon.txbiomedgenetics.org/Rhesus_Results/RhesusMapSummary.php, based on two papers (http://www.ncbi.nlm.nih.gov/pubmed/17010566, http://www.ncbi.nlm.nih.gov/pubmed/16321502,
and possibly further additional work). Markers downloaded from same source. Positions obtained for each marker by standard ePCR.

Medicago truncatula:
The map is the UMN Young 2006 integrated map, described in part here (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1456377/) and available from the Legume Information Service here: http://cmap.comparative-legumes.org/cgi-bin/cmap/map_set_info?map_set_acc=MtYoungUMinn2006.

We obtained marker information (primer sequences) from medicago.org (http://www.medicago.org/genome/downloads/Mt_genetic_markers_Oct_2009.txt) and mapped markers to the genome with ePCR as per standard approach.

Meleagris gallopavo
Map, markers, and positions downloaded from the supplement of this paper: http://www.biomedcentral.com/1471-2164/11/647

Mus musculus
The map is the new standard map, published here (http://www.genetics.org/content/182/4/1335.full), and available to download here (http://cgd.jax.org/mousemapconverter/). Coordinates were updated from NCBI Build 37 to GRCm38 coordinates using UCSC LiftOver.

Oryza sativa
The map is derived from the published map described here (http://www.ncbi.nlm.nih.gov/pubmed/23918062). To generate the map, we downloaded data and R scripts from http://www.ricediversity.org/data/ and the supplemental materials of the listed reference and recapitulated the map reconstruction algorithm in R/qtl. The map was confirmed identical to the published version. The coordinates of each marker were converted from MSU 6 to MSU 7 (IRGSP 1.0) using the coordinate conversion tool at gramene.org.

Oryzias latipes
Download mappedGeneticmarkers track from the Medaka browser, described here: http://medaka3.utgenome.org/~kobayashi/UTGBmedakaHELP/GneticMarkers.html. File is medaka_version1.0.mapping.xml. Parse markers and genetic map positions, get primer information from SNPs_markers(TUN_series).xls also available from the Genetic Markers section of Medaka browser. Map to genome by ePCR per standard approach. We only used markers and map positions derived from the HdrR-HNI backcross detailed in http://www.ncbi.nlm.nih.gov/pubmed/17554307 and http://www.ncbi.nlm.nih.gov/pubmed/16226856

Ovis aries
The domestic sheep map is version 4.7 of the SM/IMF map, available from NCBI MapViewer ftp site as 9940.SM4.7.md. This represents the latest version of the sex-averaged map originally described in http://www.ncbi.nlm.nih.gov/pubmed/11435411. Since this is exclusively an STS map, genomic positions for each marker were obtained by cross-referencing the seq_sts.md file also available from NCBI (positions are computed by ePCR at NCBI).

Papio anubis
Map downloaded from http://baboon.txbiomedgenetics.org/Bab_Polymorphisms/ChromInfoBL.php, primers downloaded from same site, mapped to genome by ePCR using standard approach. Note: mapping data is from *P. hamadryas*.

Populus trichocarpa.
Map and some marker information is derived from the supplemental materials of this paper: http://www.ncbi.nlm.nih.gov/pubmed/19220791.

Sequence information for probes comes from the published map paper; primer information for SSRs comes from http://web.ornl.gov/sci/ipgc/ssr_resource.htm. We made a few assumptions about nomenclature to link the map to the SSR file: the r before the name of some markers is meaningless, and that O = ORPM_, P = PMGC_, and W = WPMS_. Markers ending in (a) or (b) are assumed to be the same thing, and are kept only if they match the genome location.

Mapping of sequence to genome coordinates is via ePCR for primers and BLASTN for probes, using standard approach.

Prunus persica
The genetic map we are using is the T x E map, available (along with several other maps) at http://www.rosaceae.org/species/prunus_persica/genome_v1.0 along with Peach v1.0 positions for all markers (http://www.rosaceae.org/sites/default/files/peach_genome/GDR_markers_gbrowse.xls)

Prunus mume
The genetic map was extracted from http://www.ncbi.nlm.nih.gov/pubmed/23555708, which included primer sequence as supplemental material. Genomic locations were assigned with ePCR per standard approaches.

Setaria italica
The map, markers, and positions all come from Supplemental Table 4 of the map reference (http://www.nature.com/nbt/journal/v30/n6/full/nbt.2196.html).

Sorghum bicolor
The genetic map is the CIRAD map, available as supplemental information from here (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671505/). Mapped positions in the genome are obtained from the analysis published here (http://www.ncbi.nlm.nih.gov/pubmed/21484332).

Sus scrofa
Direct estimates of recombination rate between 1 Mb bins were obtained from http://www.ncbi.nlm.nih.gov/pubmed/23152986 and converted to a genetic map in cM for analysis.

Zea mays
We used the NAM map, which is available with markers already mapped from http://www.panzea.org/db/gateway?file_id=NAM_map_and_genos.