Thoughts on obligate asexuality

ruzzo

3/21/2018

To Do

is recessive vs 1/sqrt(N) modeled right?

From 11/24/17 email to julie: "There's a slight discrepancy between two simulations that should be identical (44% on top of pg 20 vs 52% top of pg 27) that I need to debug; hopefully minor." (This is now pg 23 bottom growex4(bx=0.1), 44.1% vs pg30 last plot under "hitching", growex4q(), bx=.1 52.6%)

Would be clearer if I use the intro's 2a/b/c labeling throughout, including code?

Fill in the several placeholders/questions I left in the hitchhiking section.

Plus general proofread and cleanup.

Executive Summary

We've found the L-genotype at 5 of 7 locations, and by far the most plausible explanation for the 5 is that they are asexual/clonal derivatives of a common ancestor that have become far-flung and at least a significant minority population at these locations. This note considers whether this implies that L is an *obligate* asexual, as opposed to a facultative sexual for which asexual growth has been especially effective. I consider 4 models:

1. Obligate asexual
2. Facultative sexual, in which meiosis happens less frequently than in "wild type", where this reduced frequency is:

   2a. Ablated in all sexual offspring (e.g., a "complex trait" involving several genetic loci, hence unlikely to remain intact through meiosis/fertilization)

   2b. A simple Mendelian recessive trait (i.e., a single locus, but only cells homozygous for this non-wildtype allele exhibit the reduced frequency of meiosis)

   2c. As in 2b, but a *dominant* Mendelian trait.

What I find is that the obligate asexual has an exponential growth advantage over wildtype, for the simple reason that meiosis is slower than mitosis. Under reasonable (I think) parameter choices, it will sweep the world in tens to hundreds of years, completely replacing wildtype. Reduced frequency of sexual reproduction *also* gives an exponential growth advantage, for the same reason, but the outcomes are somewhat different:

- Only model 2a results in a large, stable clonal population.
- Model 2b *transiently* shows a significant clonal population (whose size is somewhat sensitively dependant on parameter choices).
- With model 2c, the clonal population remains miniscule.

However, all three facultatively sexual models result in strong penetration of clonal haplotypes into the non-clonal population. E.g., the clonal genotype is transient in model 2b precisely because the relevant gene conferring lowered frequency of sex spreads in the non-clonal population, giving the same growth advantage to some recipient non-clones (i.e., homozygotes). Furthermore, I believe this injection into the non-clonal population would leave signatures that would be easily recognizable *if* we had a large enough sample of thaps genomes from the wild. In particular, in model 2a, SNPs initially present in the non-L population but not in L would be quickly driven to extinction, leaving behind a non-L population that looks like an HWE version

of L. Model 2b has a different fingerprint—the L-population peaks then declines, while the slow sex gene itself goes to fixation. Perhaps most surprisingly, L-clade polymorphisms *unlinked* to the slow sex gene rise in the non-L population, then stabilize, leaving about one hundred thousand SNPs at similar and substantial levels in the population, in sharp contrast to neutralist theory, which predicts that most alleles will appear at very low frequencies. Thus, in principle, these two models seem to be testable, based on collecting sufficiently many new environmental thaps samples to support or refute these predictions. Notably, our two H-clade genotypes do *not* display these signatures, which argues against facultative sex, albeit weakly since "2" isn't a very large sample. However, the existence of H-clade is *also* not consistent with the asymptotic state predicted by model 1, suggesting that our H-clade samples are either "lucky" picks at just the right time in a transient system in which they were or are headed towards extinction, or (and I think more likely) the result of effects outside these models, e.g., geographical isolation, adaptation to local conditions, and/or alternative selective advantage that is on par with the "slow sex" advantage (but probably not sexually mixing with it, else the hybrid would dominate both).

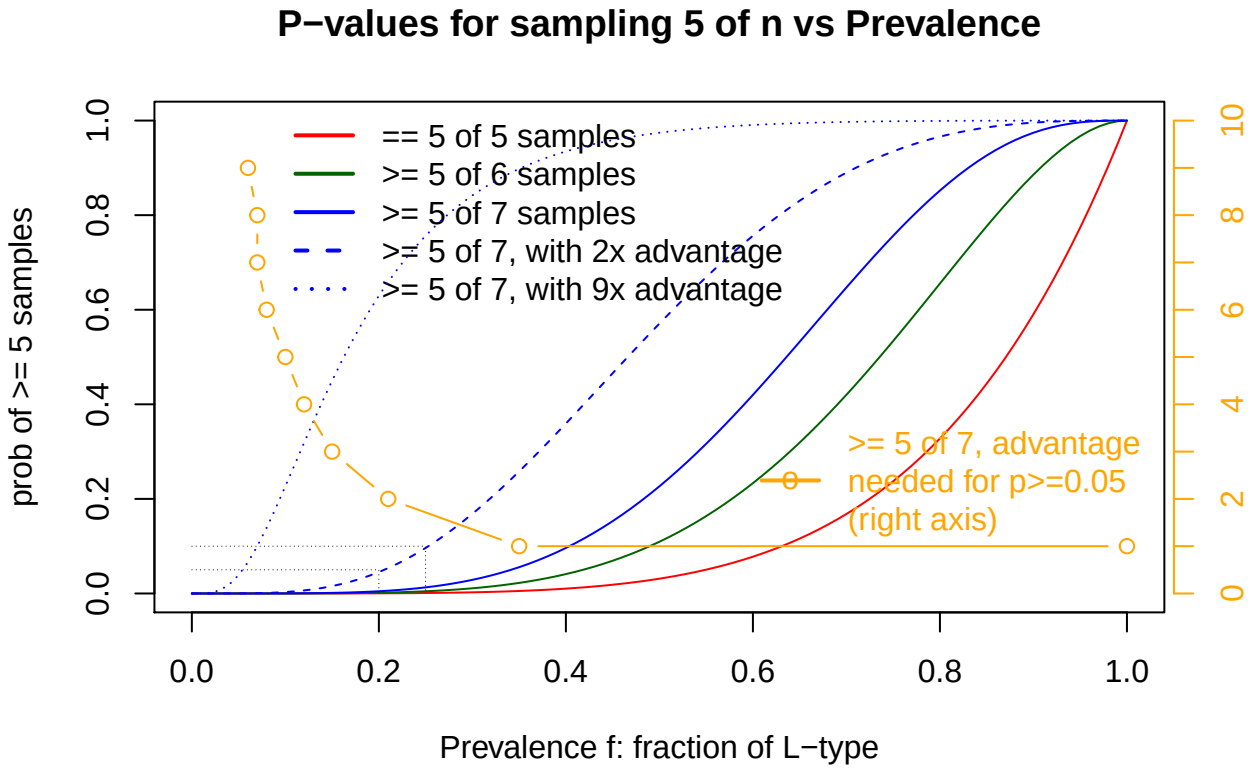
In short, 2c is inconsistent with widespread observation of L, 2b is only transiently consistent with that, and although 2a does give a stable clonal subpopulation, models 2abc all imply substantial nonclonal subpopulations having specific (but different) genetic fingerprints, none of which are consistent with either observed H population. This leads me to believe that Model 1 is most likely (and that some other effects explain the H-clade).

0. Prelude: L is at least a substantial minority

Five of seven isolates are L-clade genotype. Obviously, it must have been present at those 5 sites. Even though it may not have been the only genotype present at those sites, intuitively it must have been relatively common at each of them, otherwise we would not have isolated it 5 times. As a simple quantitative model of this, assuming it constituted a fraction f of thaps individuals at each site, and assuming that any individual was equally likely to be sampled, the probability of seeing L in 5 or more of n tries, $5 \le n \le 7$ is a simple binomial tail probability, plotted below as a function of f. (I consider $5 \le n \le 7$ since it is unclear whether we missed it twice by chance or because it is absent from one or both locales.) We cannot rule out the possibility that L is especially adept at being brought into culture, but even if so, unless that ability is many times greater than average, it is unlikely that L would have been isolated at 5 of 7 sites unless it was at least a significant minority population at these locations. E.g., under these scenarios, if the L-clade is at most a quarter of the population at each site, we have at most a 10% chance of successfully repeating the 5 of 7 sampling that we observed; lowering L's frequency to 1/5 reduces this to at most a 5% chance.

```
f  <- seq(from=0.0, to=1.0, by=0.01)     # sequence of f values to plot
s5 <- pbinom(4, 5, f, lower.tail=FALSE)  # prob of >4 successes in 5 trials
s6 <- pbinom(4, 6, f, lower.tail=FALSE)
s7 <- pbinom(4, 7, f, lower.tail=FALSE)
howmany <- 9                             # number of "s7k"'s to build.
# suppose L has a 2x advantage over others at being brought into culture.  That is
# equivalent to raising its actual proportion to 2f/((1-f)+2f) = 2f/(1+f) from f.
s72 <- pbinom(4, 7, 2*f/(1+f), lower.tail=FALSE)
plot(f, s5, type='l', col='red',
     xlab='Prevalence f: fraction of L-type',
     ylab='prob of >= 5 samples',
     main='P-values for sampling 5 of n vs Prevalence')
lines(f, s6,  col='darkgreen')
lines(f, s7,  col='blue')
lines(f, s72, col='blue',lty=2)
lines(c(0,rep(.25,2)), c(0.1, 0.1, 0), lty=3, lwd=.5)
lines(c(0,rep(.20,2)), c(.05, .05, 0), lty=3, lwd=.5)
legend(.08,1.05, bty='n',
       legend=c('== 5 of 5 samples',
```

```
                '>= 5 of 6 samples',
                '>= 5 of 7 samples',
                '>= 5 of 7, with 2x advantage',
                paste('>= 5 of 7, with ',howmany,'x advantage',sep='')),
       col=c('red','darkgreen','blue','blue', 'blue'),
       lty=c(    1,          1,     1,     2,      3), lwd=2)
# Flipping the question, as a function of f, how large a "going into culture" advantage is
# needed to be sure that we have at least a .05 chance of seeing 5 of 7?  Plotted in
# orange below, with legend & axis at right of fig.
axis(4, at=(0:10)/10, labels=c('0','','2','','4','','6','','8','','10'), col='orange',
     col.axis='orange', line=.25)
thresh <- numeric(howmany)
for(k in 1:howmany){
  s7k <- pbinom(4, 7, k*f/((1-f)+k*f), lower.tail=FALSE)
  thresh[k] <- which.max(s7k >= 0.05)
}
lines(f, s7k, col='blue', lty=3)
lines(c(1,f[thresh]), c(.1, (1:howmany)/10), type='b', col='orange')
legend(.6, .45, bty='n', legend='>= 5 of 7, advantage\nneeded for p>=0.05\n(right axis)',
       col='orange', text.col='orange', pch='o', lwd=2)
```



1. Model 1: asexual reproduction = exponential growth advantage

This is an idea that we kicked around very soon after we started considering asex as a possibility, but we never explored it quantitatively. I find the quantitative results surprising, and it may dictate changes to the main body of the paper. In short, the "global fitness" of L, and "bursting the niche" all may be no more than simple consequences of one trait: L commits to meiosis less often than H. (But Julie raised objections to this revised interpretation; *TODO: find that email, probably oct '17, and integrate a summary here...*)

Here's a simple model, with a few parameters. For concreteness, I have phrased it in terms of (I hope) plausible numerical parameter values, but the model is fully parameterized and easily evaluated with different numerical values, and I think the results do not qualitatively change as these values vary. Nevertheless, by all means let's adjust them if you think my guesses are off base. (I picked integer values, but they don't need to be.)

First, assume that, under average conditions, thaps undergoes mitosis

- $a = 180$

times per year, i.e., about every other day. This applies to both L and non-L cells. (I say "non-L" rather than "H" because it gets more complex below, and I don't need to assume that our 2 non-L samples are especially representative.) Further assume that meiosis/fertilization, when it happens, takes

- $k = 2$

times longer than the average inter-mitotic interval, i.e. ~4 days. Assume that L is strictly asexual, and that non-L is facultatively sexual, undergoing meiosis approximately

- $b_y = 0.5$

times per year, and that for each cell committing to gametogenesis, an average of

- $f = 2$

viable, fertilized offspring are produced. (The numbers below matter only insofar as to be super-concrete about the definition and interpretation of f: if 60% of meiotic cells made 3 eggs each, and the other 40% made 99 sperm each, and 80% of eggs are fertilized, then f would be $0.6 \cdot 3 \cdot 0.8 = 1.44$; i.e., sperm are irrelevant, except in so far as eggs get fertilized, but all meiotic cells count, whether they make egg or sperm. Similarly, if 50% of meiotic cells make 4 eggs, 90% of which are fertilized, then $f = 0.5 \cdot 4 \cdot 0.9 = 1.8$. For my simulations, "$f = 2$" is just a wild guess. Julie sent some nice numbers on auxospore counts and the like in the wild (10/3/2017 email "Re: asex and f"), but I don't yet see how to translate these into estimated "f".)

In this scenario, growth of L and non-L are evenly matched *except* during non-L's occasional meiotic interludes. However, the later alone gives a slight but exponential advantage to L, which of course compounds over time. To quantify this, I propose the following simple, discrete model: split one year into $a/k = 90$ *epochs*; in each epoch, L divides k times, yielding 2^k offspring. Meanwhile, a small fraction

- $\beta_y = b_y/(a/k) = 1/180$

of non-L cells commit to gametogenesis, generating, on average, f descendents each, while $1 - \beta_y = 179/180$ of non-L cells divide k times mitotically. (β_y is chosen to match non-L's overall meiotic rate: b_y meioses per year divided by a/k epochs per year equals β_y meioses per epoch.) If $f = 2^k$, both subpopulations would expand equally during each epoch, but in the more plausible scenario where $f < 2^k$, the L-subpopulation has a growth advantage.

> SIDEBAR: the pop gen book I happen to have (Gillespie "Population Genetics, a concise guide," 2nd ed 2004, Johns Hopkins Press) seems to prefer to quantify things in per-generation terms. E.g. (p63), for 2 alleles A_1, A_2, suppose the absolute viabilities, i.e., probabilities of survival, of the A_iA_j genotyppe is w_{ij}. The "selection coefficient s" for an allele A_2 vs A_1 is 1 minus its relative viability $s = 1 - w_{22}/w_{11} (\geq 0$ with the convention that A_1 is fitter, i.e. $w_{11} \geq w_{22})$. In our scenario, where "generation time" is a key trait of interest, it's not totally clear how to make an analogous definition, but I think the following is reasonably fair and on a reasonably similar scale: in one epoch, L (assumed obligate asexual) has 2^k descendants, whereas non-L averages $\beta_y * f + (1 - \beta_y) * 2^k$, so those are my "absolute viabilities", but rescale that to the k mitotic generations involved. Hence,
>
> $$\frac{\beta_y * f + (1 - \beta_y) * 2^k}{2^k} = (1 - s)^k,$$
>
> or
>
> $$s = 1 - \left(1 - \beta_y(1 - f/2^k)\right)^{1/k}$$

(which is approximately $\beta_y/4$ when $f = k = 2$).

```
sel.coeff <- function(f=2, k=2, a=180, by=0.5){
  betay <- by/(a/k)
  return(1-(1-betay*(1-f/2^k))^(1/k))
}
sel.coeff(by=c(0.2, 0.5, 1.0))
```

```
## [1] 0.000555710 0.001389855 0.002781647
```

> These strikes me as completely plausible selection coefficients; Gillespie notes in a couple of examples that 1% is a large but not crazy coefficient; `sel.coeff(`$b_y = 0.5$`)` is 7x smaller. Using a transformation like this, we could recast our parameterizations below into more conventional terms (rather than just reproductive rates), but I don't think I'll bother at this point. END SIDEBAR

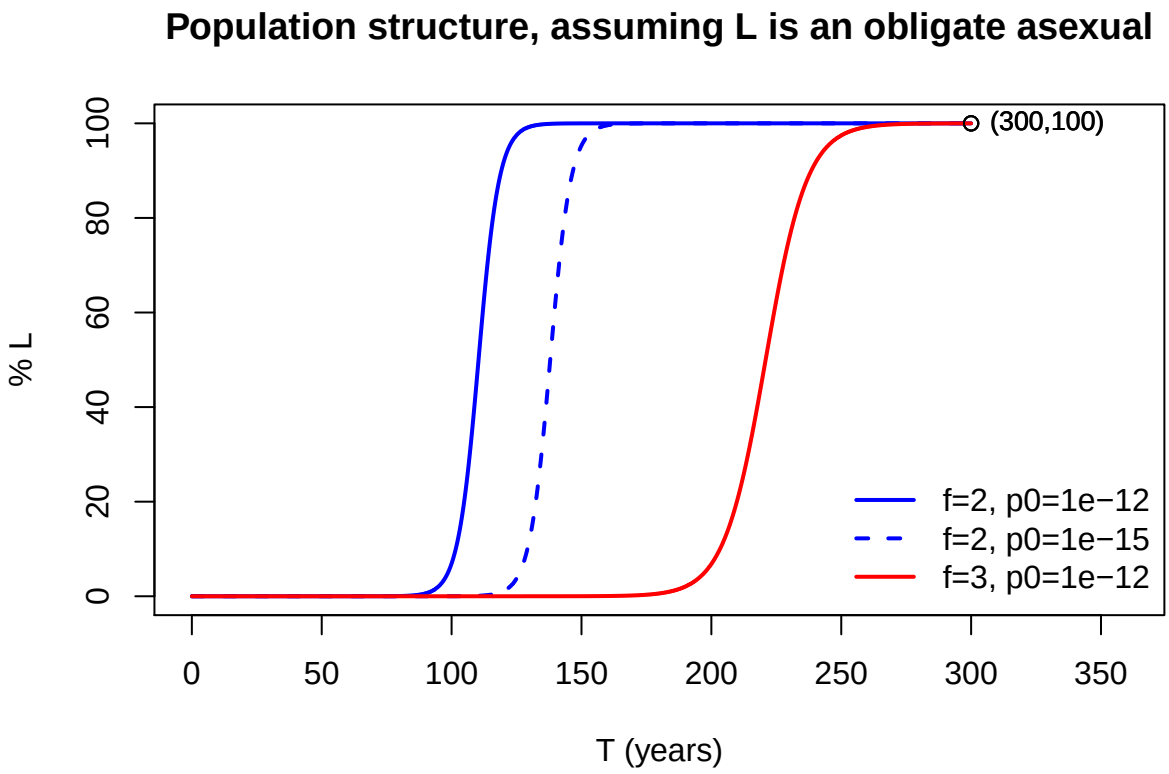
The outlined scenario is simple enough to solve analytically (I think), but subsequent ones are more challenging, so I prefer to solve each numerically in the R code below. Specifically, let x_i represents L's relative fraction of the total population at the start of the i-th epoch, and let y_i represent the non-L fraction. The function "grow1" below, for an interval spanning "yrs" years in total, returns a trio of vectors x, y, t (perhaps plus ancillary information) where x_i, y_i are as above and t_i is the time (in years) to the start of the i-th epoch. "Growth" of each sub-population follows the formulas above, then each is re-normalized to a total population of 1.0 at the end of each epoch. (Think of this as mortality affecting each sub-population equally so as to maintain a constant total population size. In particular, neither subpopulation is assumed to have a fitness advantage, other than through the difference in meiotic rates. Also, the model is fully deterministic, and in particular random genetic drift is not modeled; drift would be especially significant for rare subpopulations, sometimes leading to their extinction, sometimes not.) The parameter $p0 = 10^{-12}$ defines the initial proportion of L in the total population.

```
# grow1 simulates 'yrs' years of growth under Model 1
grow1 <- function(yrs, a=180, by=0.5, k=2, f=2, p0=1e-12){
  n <- round(yrs*a/k) # number of epochs to simulate
  x <- numeric(n)     # proportion in L at start of each epoch
  y <- numeric(n)     # proportion non-L at start of each epoch
  betay <- by/(a/k)   # meioses per epoch in non-L
  x[1] <- p0          # initial proportion in L
  y[1] <- 1-x[1]      #   and non-L
  for(i in 2:n){
    x[i] <- x[i-1]*2^k                          # L grows mitotically
    y[i] <- y[i-1]*((1-betay)*2^k + betay*f)    # non-L grows both ways
    x[i] <- x[i]/(x[i]+y[i])                    # re-normalize
    y[i] <- 1-x[i]                              # re-normalize
  }
  t <- (0:(n-1))/a*k  # rescale epoch numbers to years
  return(list(x=x, y=y, t=t))
}
```

The three sample simulations below show that, with the default parameters I defined above, L will expand from 1 cell per trillion to essentially 100% of the population in about one century (solid blue curve). Lowering the initial proportion of L by three orders of magnitude delays the transition by 2–3 decades (dashed blue curve). Similarly, raising f to 3 from 2 (in comparison to $2^k = 4$) delays the transition by about a century (red curve). The bottom line is the same in all cases: all else being equal, strictly asexual growth is sufficient to dominate the population in a few hundred years. It is reassuring that the time scale for asexual dominance is comparable to our age estimate for the emergence of the L-genotype, but honestly I think it is a mistake to over-interpret that. Both include enough simplifying assumptions and "ball park" estimates that we should assume they come with wide error bars. Nevertheless, it is encouraging that they are not wildly discordant.

Feel free to experiment with the simulation parameters.

```
# annotate a specific x,y point
showxy <- function(x, y, dy=0, show.point=TRUE){
  if(show.point){
    points(x,y)
  }
  text(x, y+dy, paste('(', round(x), ',', round(y,1), ')', sep=''), cex=0.8, pos=4)
}

# plot1 plots a growth series g (as returned by grow1) in specified color, line type
plot1 <- function(g,col,lty=1){
  n <- length(g$x)
  p <- g$x * 100                          # rescale x to percent
  lines(g$t, p, lwd=2, col=col, lty=lty)  # plot x vs t
  showxy(g$t[n], p[n])                    # annotate last point in series
}

# plot a few examples
growex1 <- function(years=300, a=180, by=1/2, k=2, f=2, p0=1e-12){
  # Set up plot axes
  plot(NULL,type='n', xlim=c(0,1.2*years), ylim=c(0,100), xlab='T (years)', ylab='% L',
       main='Population structure, assuming L is an obligate asexual')
  asexf2  <- grow1(years,a,by,k,f,  p0)       # simulation w/ default params
  asexf2a <- grow1(years,a,by,k,f,  p0/1000)  # sim w/ lower initial proportion
  asexf3  <- grow1(years,a,by,k,f+1,p0)       # sim w/ more sexual offspring
  plot1(asexf2,  "blue")
  plot1(asexf2a, "blue", lty=2)
  plot1(asexf3,  "red")
  legend('bottomright',
         legend=c('f=2, p0=1e-12','f=2, p0=1e-15', 'f=3, p0=1e-12'),
         col=   c('blue',         'blue',          'red'),
         lty=   c(1,               2,               1),
         lwd=2, bty='n')
}
growex1()
```



[A tangent on p0: Here's a very crude estimate of global thaps population. Assuming phytoplankton are largely confined to the top 10 meters of the ocean (a volume of 3.4×10^{18} liters), and an average density of 10 cells per liter (you said "thousands" per liter in bloom conditions), gives 3×10^{19} cells:

```
r   <- 6371000              # radius of earth (m)
A   <- 2/3*4*pi*r^2          # area of oceans (m^2)
V   <- A * 10 * 100^3/1000   # liters of sea water within 10m of surface
cpl <- 10                    # cells per liter (non-bloom)
cat(">> Est.", signif(10*V,1), 'cells, assuming global avg of', cpl,
    'cells per L (non-bloom) x', signif(V,2), 'liters.\n')
```

```
## >> Est. 3e+19 cells, assuming global avg of 10 cells per L (non-bloom) x 3.4e+18 liters.
```

By this estimate, perhaps I should set $p0 \approx 10^{-19}$. Again, this could be way off and please correct if so, but I don't think it will fundamentally change the picture—as noted in the first simulation, dropping the initial proportion a thousand fold only adds a few decades to the time line, and I think the early stages of growth from a single ancestral L cell will be dominated by stochastic effects, anyway, i.e., my growth model above is most appropriate after L has reached a scale, say millions of cells, where stochastic effects are minor.]

2. Model 2: *Slowed* sexual reproduction *also* = exponential growth advantage

Instead of obligate asexuality, what if L just had a reduced rate of meiosis? E.g., $b_x = 0.2$ in comparison to non-L's $b_y = 0.5$? The result is almost the same: L will grow to dominate the global population in only a few hundred years. There is, however, one important twist—in this model, in the limit, L's proportion of the population is *not* 100%, for the reason that the large L population is continually creating sexual offspring that are recognizably non-L (as we've discussed before). But if we assume for now (but see next section) that these non-L cells undergo meiosis at the non-L rate b_y, then asexual growth in L holds the clonal lineage at a high level.

Modeling here is similar to that outlined in the previous section. The key difference is that in each epoch

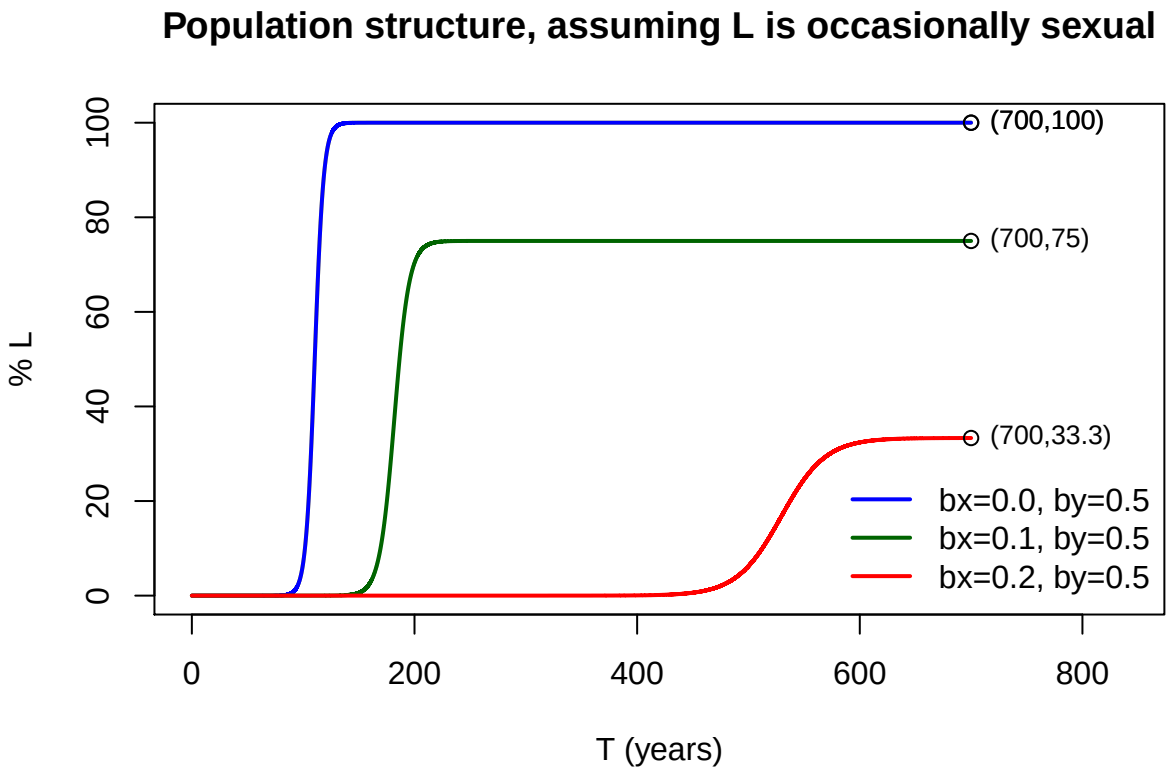
we assume that a fraction $\beta_x = b_x/(a/k)$ of the L population undergoes meiosis, while $(1 - \beta_x)$ divides mitotically k times in one epoch, as before. Also, the meiotic offspring of L are added to the *non-L* proportion of the population; this change has little effect while the L proportion is small, but serves to keep the non-L proportion from shrinking to zero after L becomes dominant. Again we are assuming both the L and non-L populations are on average adequately characterized by the parameters a, k, f, b_x, and b_y, which implies no fitness differences other than the different meiotic rates.

```
# grow2 simulates 'yrs' years of growth in the 2-component model where L & non-L
# have equal mitotic but unequal meiotic rates.
grow2 <- function(yrs, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12){
  n <- round(yrs*a/k) # number of epochs to simulate
  x <- numeric(n)     # proportion in L
  y <- numeric(n)     # proportion non-L
  betax <- bx/(a/k)   # meioses per epoch in L
  betay <- by/(a/k)   # meioses per epoch in non-L
  x[1] <- p0          # initial proportion in L
  y[1] <- 1-x[1]
  for(i in 2:n){
    # L grows mitotically:
    x[i] <- x[i-1]*(1-betax)*2^k
    # non-L grows both ways *and* absorbs meiotic offspring of L
    y[i] <- x[i-1]*betax*f + y[i-1]*((1-betay)*2^k + betay*f)
    x[i] <- x[i]/(x[i]+y[i])  # re-normalize
    y[i] <- 1-x[i]            # re-normalize
  }
  t <- (0:(n-1))/a*k  # rescale epoch numbers to years
  return(list(x=x, y=y, t=t))
}
```

The three examples shown below illustrate hypothetical meiosis in L never, every 10 years and every 5 years, all in comparison to every 2 years in non-L. As seen in the graphs below, reducing the difference in meiotic rate between L and non-L delays L's transition and lowers its asymptotic proportion, but the general trends are qualitatively the same—L rises from 1 part per trillion to a large fraction of the population in a few centuries. (A fourth, yellow, curve was also plotted, based on the "grow1" asexual simulation given earlier, but this curve is *completely* overlaid by the blue $b_x = 0$ curve from "grow2". This is expected, since the second model with $b_x = 0$ is logically equivalent to the first model; I included it as a sanity check.)

```
# plot a few examples
growex2 <- function(yrs=700, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12){
  # Set up plot axes
  plot(NULL,type='n', xlim=c(0,1.2*yrs), ylim=c(0,100), xlab='T (years)', ylab='% L',
       main='Population structure, assuming L is occasionally sexual')
  asex  <- grow1(yrs, a,         by, k, f, p0) # *a*sex simulation w/ default params
  sex00 <- grow2(yrs, a, bx=0.0, by, k, f, p0) # sanity check: new w/ bx=0 should = old
  sex01 <- grow2(yrs, a, bx=0.1, by, k, f, p0) # sim w/ more sexual offspring
  sex02 <- grow2(yrs, a, bx=0.2, by, k, f, p0) # sim w/ even more
  plot1(asex,  'yellow')
  plot1(sex00, 'blue')
  plot1(sex01, 'darkgreen')
  plot1(sex02, 'red')
  legend('bottomright',
         legend=c('bx=0.0, by=0.5', 'bx=0.1, by=0.5', 'bx=0.2, by=0.5'),
         col=   c('blue',           'darkgreen',      'red'),
         lwd=2, bty='n')
}
```

```
growex2()
```



(Once L becomes dominant in the population, its meiotic offspring frequently will be consanguineous, including direct L cross L matings. Inbreeding-depression might reduce the non-L proportion of the population, but this effect seems insufficiently strong to eliminate the non-L population, especially given that we believe L itself to be inbred.)

So, superficially, this model is plausibly consistent with our observations *without* mandating that L is an obligate asexual; it just requires that L undergoes meoisis less frequently (coupled with meiosis being less "productive" than mitosis, due to longer cell division times, incomplete fertilization, etc., all of which are expected). However, in these models the non-L populations largely will be descendants of L, so it is not clear that they could maintain the large pool of non-L SNPs that we see in both H isolates. This is more fully explored in section 3.

Furthermore, after allowing for what I think is the most plausible genetic explanation for reduced meiotic frequency, I think obligate asexual is the most likely scenario: see section 4.

3. Model 2-prime: Tracking L's offspring

In the "slow sex" model model considered in the previous section, the sexual offspring of L carry L-alleles into the non-L population. The extent of this introgression is explored in this section. Somewhat surprisingly, this happend to an extreme extent, sufficiently so that the stable L/non-L coexistance depicted above cannot explain our observed L/H snapshot. The simulations below first consider only non-L cells that can trace all ancestors to L (e.g., a direct L-L cross, or crosses between such crosses). The second analysis drops this "pure-L-ancestry" restriction. In retrospect, the second analysis is the more interesting, but I'll leave the first just because I thought of it first...

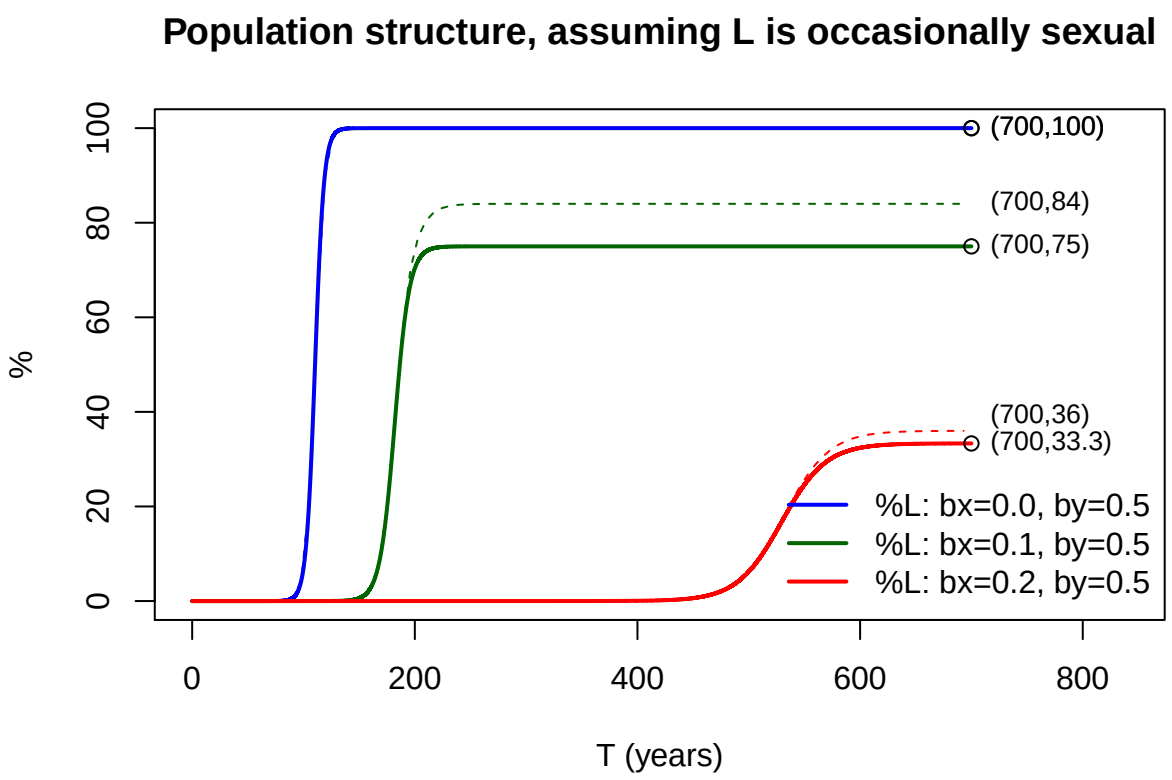
```
# grow2 simulates 'yrs' years of growth in the 2-component model where L & non-L have
# equal mitotic but inequal meiotic rates.  Additionally, track the proportion of the
# population that has purely L-ancestry (i.e., direct result of L-L crosses, of L crossing
```

```
# with them and with each other, etc.).  The later are assumed to have the same meiotic
# and mitotic rates as other non-L genotype, but would be genetically recognizable due to
# their near-absence of non-reference nucleotides not seen in L. Plot later as dashed line
# atop L portion.
grow2pold <- function(yrs, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12){
  n  <- round(yrs*a/k) # number of epochs to simulate
  x  <- numeric(n)     # proportion in L
  z  <- numeric(n)     # proportion in LxL
  y  <- numeric(n)     # proportion other non-L
  betax <- bx/(a/k)    # meioses per epoch in L
  betay <- by/(a/k)    # meioses per epoch in non-L
  x[1]  <- p0          # initial proportion in L
  z[1]  <- 0
  y[1]  <- 1-x[1]
  for(i in 2:n){
    # total meiotic fraction
    m <- x[i-1]*betax + (z[i-1]+y[i-1])*betay
    # fraction of them that are purely-L-ancestory
    p <- (x[i-1]*betax + z[i-1]*betay)/m
    q <- 1-p
    # L grows mitotically:
    x[i]  <-  x[i-1]*(1-betax)*2^k
    # non-L grows both ways *and* absorbs meiotic offspring
    # of L, with p^2 fraction retaining L-purity
    z[i]  <- z[i-1]*(1-betay)*2^k +    p^2  * m * f
    y[i]  <- y[i-1]*(1-betay)*2^k + (1-p^2) * m * f
    w <- x[i]+z[i]+y[i]
    x[i]  <-  x[i]/w    # re-normalize
    z[i]  <-  z[i]/w    # re-normalize
    y[i]  <-  y[i]/w    # re-normalize
  }
  t <- ((1:n)-1)/a*k  # rescale epoch numbers to years
  return(list(x=x, z=z, y=y, t=t))
}
```

```
# plot2p plots a growth series g (as returned by grow2p) in specified color, line type
plot2pold <- function(g,col,lty=1, label=F){
  n <- length(g$x)
  p <- g$x * 100                          # rescale x to percent
  lines(g$t, p, lwd=2, col=col, lty=lty)  # plot x vs t
  showxy(g$t[n], p[n])                    # annotate last point in series
  # now plot growth of z atop x as a thinner dashed line. (plot subseq of about 100
  # points,  otherwise dashing is obscured)
  ss <- seq(from=1, to=n, by=ceiling(n/100))
  lines(g$t[ss], (g$x[ss]+g$z[ss])*100, lwd=1, col=col, lty='dashed')
  if(label){ # label last point in this one too?
    pp <- (g$x[n]+g$z[n])*100
    dy <- ifelse(pp < 50, 3, 0) # hack to avoid label collision for small pp
    showxy(g$t[n],pp, dy=dy, show.point=FALSE)
  }
}
```

```
# plot a few examples
growex2pold <- function(yrs=700, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12){
```

```
  # Set up plot axes
  plot(NULL,type='n', xlim=c(0,1.2*yrs), ylim=c(0,100), xlab='T (years)', ylab='%',
       main='Population structure, assuming L is occasionally sexual')
  asex  <- grow1(yrs, a,         by, k, f, p0) # *a*sex simulation w/ default params
  sex00 <- grow2pold(yrs, a, bx=0.0, by, k, f, p0) # sanity check: new+(bx=0) should = old
  sex01 <- grow2pold(yrs, a, bx=0.1, by, k, f, p0) # sim w/ more sexual offspring
  sex02 <- grow2pold(yrs, a, bx=0.2, by, k, f, p0) # sim w/ even more
  #sex022<- grow2pold(yrs, a, bx=0.22, by, k, f, p0) # sim w/ even more
  plot1(asex,  'yellow')
  plot2pold(sex00, 'blue')
  plot2pold(sex01, 'darkgreen',label=T)
  plot2pold(sex02, 'red', label=T)
  legend('bottomright',
         legend=c('%L: bx=0.0, by=0.5', '%L: bx=0.1, by=0.5', '%L: bx=0.2, by=0.5'),
         col=   c('blue',               'darkgreen',          'red'),
         lwd=2, bty='n')
}
growex2pold()
```



Now look at overall allele frequencies in the whole non-L population.

```
# In addition to the functionality of grow2pold above, this version tracks the allele
# frequency, in the non-L-, non-pure-L-population, of any allele that is
# identical-by-descent to an allele in L.  E.g., at a position that is polymorphic across
# the initial L/non-L pop, does growth of L also push the L-alleles out into the non-L
# population?  The reason to focus on identical-by-descent is that I can't estimate freq
# of these alleles in the initial non-L pop, but finding IBD freq over time gives a lower
# bound on overall freq. I do this separately for het and hom positions in L, i.e. initial
# freq in L are exactly 1/2 or exactly 1.
#
```

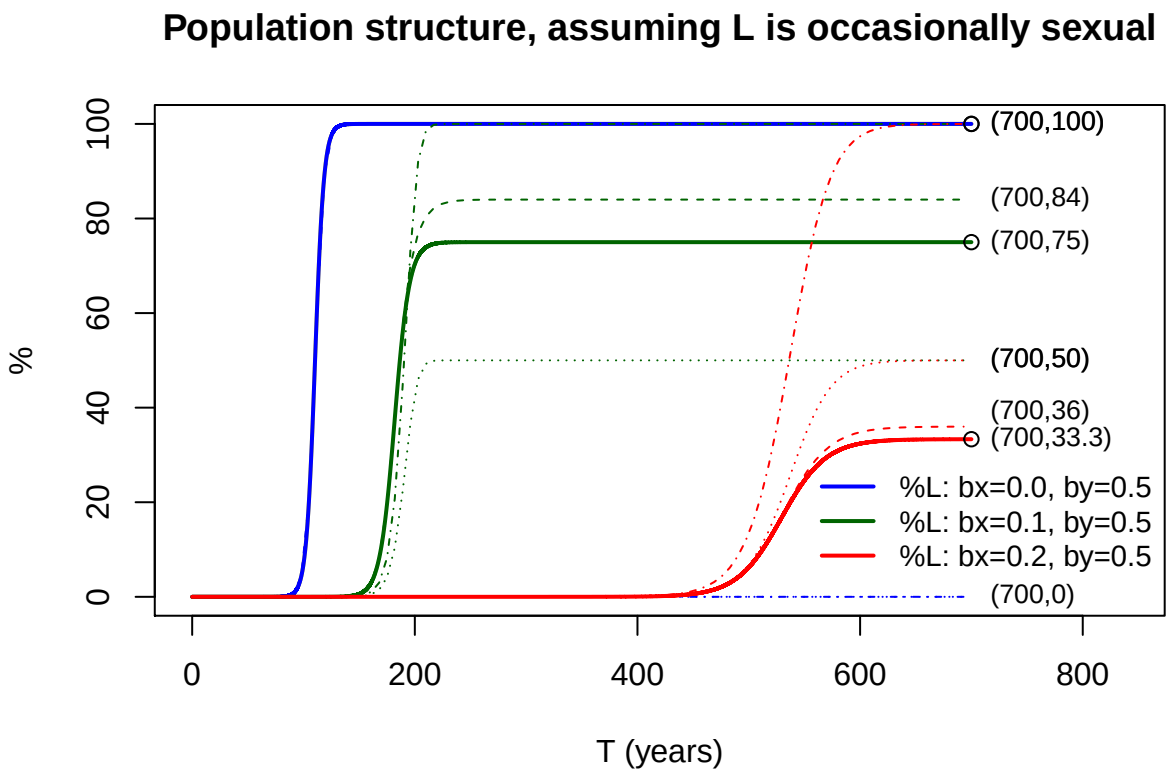
```
# The variables "x, y, z", below, record (over time) the proportion of the population in
# one of the following 3 genotype classes:
#
#   x: the L-genotype
#   z: non-L genotypes all of whose ancestors are traceable to L; e.g. (L x L) x L
#   y: all other non-L (e.g., original wildtypes and L-outcrosses)
#
# Note that in x, all positions are either homozygous for the same nucleotide across all
# cells, or heterozygous for the same pair of nucleotides across all cells.  In z, we will
# see the same allele frequencies *on average*, but the 50:50 positions will exist in HWE
# --- 25% homozygous for one allele, 25% hom for the other and 50% het.  In y, which is
# initially all but p0, we could see any allele frequencies.  A question of interest, in
# the case that x grows, is whether y can retain unique allele frequencies or whether
# outcrossing of x and z with y will push y's allele frequency structure to become more
# L-like.  This simulation suggests that yes, this happens.
grow2p <- function(yrs, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12){
  n   <- round(yrs*a/k)# number of epochs to simulate
  x   <- numeric(n)    # proportion in L
  z   <- numeric(n)    # proportion in LxL
  y   <- numeric(n)    # proportion other non-L
  y05 <- numeric(n)    # allele freq IBD-from-L in non-L pop for a het position in L
  y10 <- numeric(n)    # allele freq IBD-from-L in non-L pop for a hom position in L
  betax <- bx/(a/k)    # meioses per epoch in L
  betay <- by/(a/k)    # meioses per epoch in non-L
  x[1]  <- p0          # initial proportion in L
  z[1]  <- 0
  y[1]  <- 1-x[1]
  y05[1] <- 0
  y10[1] <- 0
  for(i in 2:n){
    # total meiotic fraction
    m <- x[i-1]*betax + (z[i-1]+y[i-1])*betay
    # fraction of them that are purely-L-ancestory
    p <- (x[i-1]*betax + z[i-1]*betay)/m
    q <- 1-p
    # L grows only mitotically:
    x[i]  <-  x[i-1]*(1-betax)*2^k
    # non-L grows both ways *and* absorbs meiotic offspring
    # of L, with p^2 fraction retaining L-purity
    z[i] <- z[i-1]*(1-betay)*2^k +    p^2  * m * f
    y[i] <- y[i-1]*(1-betay)*2^k + (1-p^2) * m * f
    # Tracking L's introgression into y's allele frequency structure: Consider one epoch.
    # Picture x,y,z as cell *counts* in the population (in units of the total pop size if
    # you prefer). Count chromosomes (2 per cell, of course) being added to y's gene pool
    # as a result of both mitosis and meiosis of y cells, and meiosis of x and z cells.
    # (Mitosis within x and z don't contribute to y.)  Oh, rats -- I really should have
    # said "count physical copies of one haploid genome, of which there are 2 per cell",
    # but I don't want to change many variable names and comments at this late stage, so
    # just pretend that "ch" is somehow an abbreviation for "copies, haplo" or somehing,
    # and that "chromosome" in many of the comments below really means "haplocopy".
    ch.y.mito <- y[i-1] * (1-betay) * 2^k * 2
    ch.y.meio <- y[i-1] * betay * f * 2
    ch.z.meio <- z[i-1] * betay * f * 2
```

```
    ch.x.meio <- x[i-1] * betax * f * 2
    # New freq in y at L's het sites: Initially, y dominates the population, and nearly
    # all chromosomes in the y gene pool in one epoch come from meiosis and, especially,
    # mitosis of y cells.  Whatever past introgression of L alleles has occurred is
    # measured by y05[] and y10[], and these (initially tiny fractions) are
    # proportionately carried forward into the next epoch.  In contrast, chromosomes being
    # added to y that originate from meioses in x or z carry L alleles at L frequencies,
    # i.e. 1/2 or 1.  The y05 and y10 values for the next epoch are simply a weighted
    # average of these contributions.  Although the number of chromosomes contributed by
    # x/z is initially small, these allele frequecies are large and seem to inexorably
    # build in the population, especially as x/z grow.
    y05[i] <- ((ch.y.mito + ch.y.meio) * y05[i-1] + (ch.z.meio + ch.x.meio) * 0.5) /
              ( ch.y.mito + ch.y.meio             +  ch.z.meio + ch.x.meio       )
    y10[i] <- ((ch.y.mito + ch.y.meio) * y10[i-1] + (ch.z.meio + ch.x.meio) * 1.0) /
              ( ch.y.mito + ch.y.meio             +  ch.z.meio + ch.x.meio       )
    w <- x[i]+z[i]+y[i]
    x[i] <- x[i]/w    # re-normalize
    z[i] <- z[i]/w    # re-normalize
    y[i] <- y[i]/w    # re-normalize
  }
  # in retrospect, no need to compute y10; it's always twice y05.  Easily seen by
  # induction on i at defining calculation above or by:
  if(max(abs(y10 - 2*y05))>1e-12){
    print(summary(y10-2*y05))
  }
  t <- ((1:n)-1)/a*k  # rescale epoch numbers to years
  return(list(x=x, z=z, y=y, y05=y05, y10=y10, t=t))
}
```

```
# plot2p plots a growth series g (as returned by grow2p) in specified color, line type
plot2p <- function(g,col,lty=1, label=F){
  n <- length(g$x)
  p <- g$x * 100                          # rescale x to percent
  lines(g$t, p, lwd=2, col=col, lty=lty)  # plot x vs t
  showxy(g$t[n], p[n])                    # annotate last point in series
  # now plot growth of z atop x as a thinner dashed line, and y05, y10 as dotted,
  # dash-dotted, resp. (plot subseq of about 100 points for these, otherwise dash/dot is
  # obscured)
  ss <- seq(from=1, to=n, by=ceiling(n/100))
  lines(g$t[ss], (g$x[ss]+g$z[ss])*100, lwd=1, col=col, lty='dashed')
  lines(g$t[ss], (g$y10[ss])*100, lwd=1, col=col, lty='dotdash')
  lines(g$t[ss], (g$y05[ss])*100, lwd=1, col=col, lty='dotted')
  showxy(g$t[n], (g$y05[n])*100, show.point=FALSE)
  if(label){ # label last point in this one too?
    pp <- (g$x[n]+g$z[n])*100
    dy <- ifelse(pp < 50, 3, 0) # hack to avoid label collision for small pp
    showxy(g$t[n],pp, dy=dy, show.point=FALSE)
  }
}
```

```
# plot a few examples
growex2p <- function(yrs=700, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12){
  # Set up plot axes
  plot(NULL,type='n', xlim=c(0,1.2*yrs), ylim=c(0,100), xlab='T (years)', ylab='%',
```

```
       main='Population structure, assuming L is occasionally sexual')
  asex  <- grow1 (yrs, a,         by, k, f, p0) # *a*sex simulation w/ default params
  sex00 <- grow2p(yrs, a, bx=0.0, by, k, f, p0) # sanity check: new w/ bx=0 should = old
  sex01 <- grow2p(yrs, a, bx=0.1, by, k, f, p0) # sim w/ more sexual offspring
  sex02 <- grow2p(yrs, a, bx=0.2, by, k, f, p0) # sim w/ even more
  #sex022<- grow2p(yrs, a, bx=0.22, by, k, f, p0) # sim w/ even more
  plot1(asex,  'yellow')
  plot2p(sex00, 'blue')
  plot2p(sex01, 'darkgreen',label=T)
  plot2p(sex02, 'red', label=T)
  legend('bottomright', bty='n', inset=c(0,.05), cex=0.9,
         legend=c('%L: bx=0.0, by=0.5', '%L: bx=0.1, by=0.5', '%L: bx=0.2, by=0.5'),
         col=   c('blue',               'darkgreen',          'red'),
         lwd=   2)
}
growex2p()
```



The legend in the figure is incomplete; here's the story. The 3 solid colored lines show the percent of the L-genotype in the population as a function of time: L sweeps the population in about a century if it's an obligate asexual (blue), and stabilizes somewhat more slowly at 3/4 or 1/3 of the population in the other two cases (green, red, resp.) Additionally, the "pure L ancestry" class rises to small but stable fraction of the population - 9% and 2.66% of the total in the 2 examples plotted above (the red and green dashed lines).

HOWEVER, asex drives allele frequencies *in the population as a whole* to match L's *individual* frequencies, *even when the L-genotype population remains a minority*, and *no matter what the "wildtype" allele frequencies are initially.* I.e., heterozygous positions in L give rise to 50% allele frequencies in the entire population (dotted green & red lines), and homozygous positions in L go to fixation (green & red dot-dashed lines), no matter what alleles preexisted in the wildtype population.

At first blush, the solid green curve above (say), looks a bit like our data, with non-L genotypes ("probably H,

right?") co-existing with L. But the foregoing analysis says that the stable non-L fraction of the population would look very different from H. Specifically, it would have *no* private SNPs, and would exhibit HWE proportions at *all* of L's SNPs.

Furthermore, we should re-visit the assumption that slow sex is a complex trait driven by the happy confluence of many SNPs in many genes/regulatory regions and hence is easily destroyed by recombination, never to reappear. Suppose that quantitatively this involves, e.g., 30 SNPs on 5 different chromosomes. Then a random genome drawn from the above pool has one chance in 2^{30} of having the same 30 heterozygous positions, and one chance in 2^{25} of having them phased the same way as in L (the 1st SNP on each chromosome is arbitrarily phased; the remaining 25 have 1 chance in 2 to be phased relative to the first in the same way as in L). (Phasing may not matter, but conservatively, assume it does.) Hence, one cell in 2^{55} *within y* will recapitulate the slow sex phenotype. While this is a miniscule fraction, given that my ballpark estimate of thaps population size is $10^{19} \approx 2^{63}$, it is virtually certain that the slow sex phenotype would re-emerge on many non-L backgrounds, which would strongly constrain the fraction of L-genotype in the population (see model 4). I didn't try to simulate this one numerically, because 30/5 are just made up numbers, and I think the case is clear enough without this, but just FYI...

In summary, assuming that L remains facultatively sexual, but at a (nontrivially) lower frequency that wildtype, and further assuming that this reduced sexual frequency phenotype is easily disrupted by recombination/fertilization (e.g., is a "complex trait" involving many genes/SNPs), then the pure-L genotype will sweep to a high fraction of the population in only a few hundred years and saturate the remaining population with its alleles just as quickly.

As another sanity check, here's a different derivation of the same result that doesn't bother to seperately track the "purely-L-ancestry" component ("z" above).

```
# Like grow2p above, this version tracks, within the non-L-population, the allele
# frequencies of nucleotides seen in L.  E.g., at a position that is polymorphic across
# the initial L/non-L pop, does growth of L also push the L-alleles out into the non-L
# population?  (Unlike the previous version, this does not separately track
# "purely-L-ancestry".) The variables "x, y", below, record (over time) the proportion of
# the population in one of the following 2 genotype classes:
#
#   x: the L-genotype
#   y: all non-L (e.g., original wildtypes, L-inbreds and L-outcrosses)
#
# Note that in x, all positions are either homozygous for the same nucleotide across all
# cells, or heterozygous for the same pair of nucleotides across all cells.  In y, which
# is initially all but p0, we could see any allele frequencies.  A question of interest,
# in the case that x grows, is whether y can retain unique allele frequencies or whether
# mating of x with itself or with y will push y's allele frequency structure to become
# more L-like.  This simulation suggests that yes, this happens.
#
# All but last parameter are as in earlier simulations; see below for "lag"
grow2q <- function(yrs, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12, lag=FALSE){
  n   <- round(yrs*a/k)# number of epochs to simulate
  x   <- numeric(n)    # proportion in L
  y   <- numeric(n)    # proportion non-L
  r   <- numeric(n)    # allele freq of an L SNP in non-L
  betax <- bx/(a/k)    # meioses per epoch in L
  betay <- by/(a/k)    # meioses per epoch in non-L
  x[1]  <- p0          # initial proportion in L
  y[1]  <- 1-x[1]
  r[1]  <- 0
  r.lag <- 0           # see 'lag' code below
  for(i in 2:n){
```

```
    # Tracking L's introgression into y's allele frequency structure: Consider one epoch.
    # Picture x,y as cell *counts* in the population (in units of the total pop size to be
    # precise). Count chromosomes (2 per cell, of course) being added to y's gene pool as
    # a result of both mitosis and meiosis of y cells, and meiosis of x cells. Mitosis
    # within x contributes to x, not y.
    ch.x.mito <- x[i-1] * (1-betax) * 2^k * 2
    ch.y.mito <- y[i-1] * (1-betay) * 2^k * 2
    ch.x.meio <- x[i-1] * betax * f * 2
    ch.y.meio <- y[i-1] * betay * f * 2
    # The proportion of x in the population at the start of the next epoch is just the
    # number of x chromosomes produced by x-meiosis, normalized by the total number of
    # chromosomes:
    ch <- c(ch.x.mito, ch.y.mito, ch.x.meio, ch.y.meio)
    x[i] <- ch.x.mito/sum(ch)
    y[i] <- 1 - x[i]
    # The new allele freq in y (for an allele that is het in L): Initially, y dominates
    # the population, and all chromosomes in the y gene pool in one epoch come from
    # mitoses of y cells and meioses of either x or y.  Whatever past introgression of L
    # alleles has occurred is measured by r[], and this (initially tiny fraction) is
    # proportionately carried forward into the next epoch. In contrast, chromosomes being
    # added to y via meioses in x carry L alleles at L frequencies, i.e. 1/2 for het sites
    # in L.  The r value for the next epoch is simply a weighted average of these
    # contributions.  (Computationally this is the rather terse expression in the next
    # line - the dot product of the vector "ch" containing the 4 chromosome counts with
    # the weight vector "(0, r, 1/5, r)", divided by the unweighted sum of the later 3
    # counts.)  Although the number of chromosomes contributed by x is initially small,
    # these allele frequecies are large and inexorably build in the population, especially
    # as x grows - each epoch adds a new pulse of L alleles to the non-L population,
    # eventually washing out any remnant of the initial non-L population.  I have not
    # separately tracked the allele frequency of sites that are initially homozygous in L.
    # An analogous formula applies, but with "0.5" replaced by "1.0".  The result is that
    # those sites always have twice the allele frequency calculated below (can be shown
    # formally by induction on i), again meaning that in the limit the non-L population
    # will be homozygous for the L allele as well.
    if(!lag){
      r[i] <- (ch %*% c(0, r[i-1], 0.5, r[i-1])) / sum(ch[2:4])
    } else {
      # Cells in y had their genomes fixed at their last meiosis, and so we should
      # probably use r[then], rather than r[i-1], however, "then" varies cell to cell.  In
      # our simple model, distribution of time (in "epochs") to last meiosis is geometric
      # with parameter betay, so find Expectation of r based on that: E[r] =
      # sum_{0<=k<i-1}p^kqr[i-k], where 1-q=p=betay. Since this turns out to make very
      # little difference, I won't push harder at it, but seemed worth checking.
      if(FALSE){
        r.lag <- sum(betay^((i-2):0)*r[1:i-1])*(1-betay)
      } else {
        # straightforward code above is quite slow, ~ 1 sec per simulated year.  Iterative
        # version below (initialized to 0) computes the same thing faster.
        r.lag <- r.lag * betay + r[i-1] * (1-betay)
      }
      # if(i %% 180 == 0){cat(i/90, r.lag, '\n')} # Debug
      #
      # I initially had tried r[E(lag)], rather than E(r[lag]) but this looked inaccurate:
```

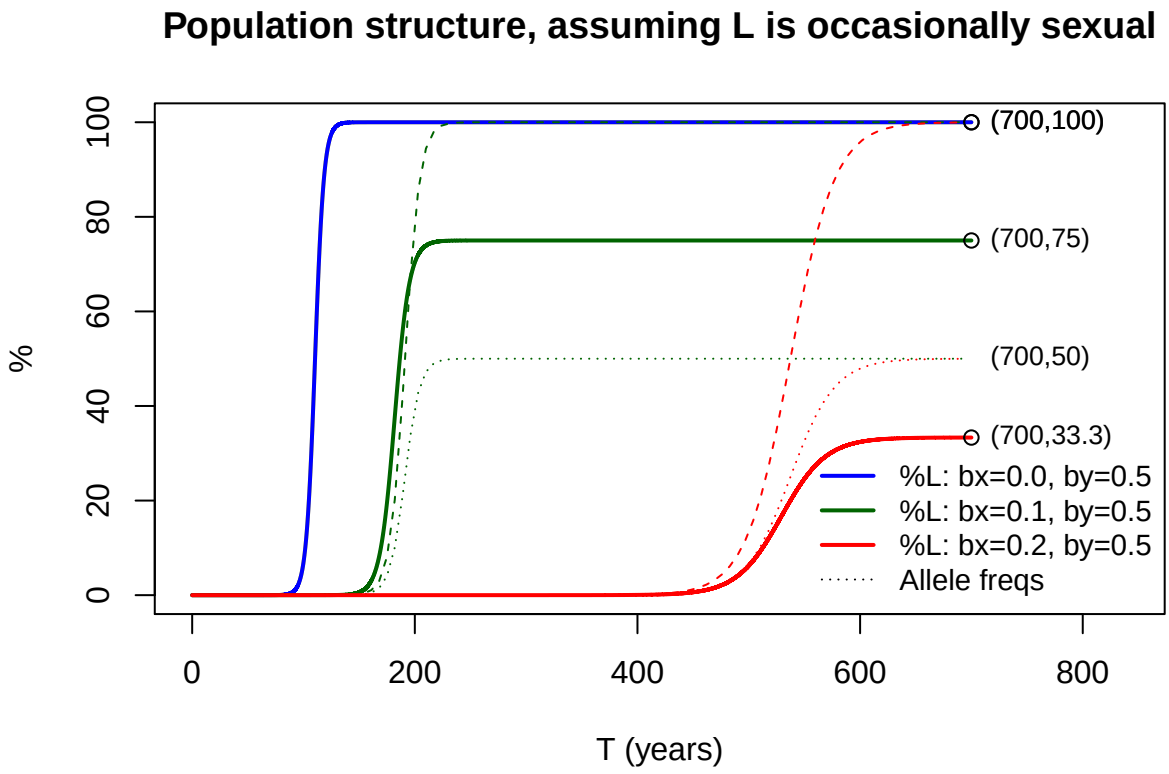
```
      #   lag.epochs <- min(i-1, max(1, 1/betay))
      #   r.lag <- r[i-lag.epochs]
      r[i] <- (ch %*% c(0, r.lag, 0.5, r.lag)) / sum(ch[2:4])
    }
  }
  t <- ((1:n)-1)/a*k  # rescale epoch numbers to years
  return(list(x=x, y=y, r=r, t=t))
}
```

```
# plot2p plots a growth series g (as returned by grow2p) in specified color, line type
plot2q <- function(g,col,lty=1, label=F, allele=T){
  n <- length(g$x)
  p <- g$x * 100                          # rescale x to percent
  lines(g$t, p, lwd=2, col=col, lty=lty)  # plot x vs t
  showxy(g$t[n], p[n])                    # annotate last point in series
  if(allele){
    # now plot growth of z atop x as a thinner dashed line, and y05, y10 as dotted,
    # dash-dotted, resp. (plot subseq of about 100 points for these, otherwise dash/dot is
    # obscured)
    ss <- seq(from=1, to=n, by=ceiling(n/100))
    lines(g$t[ss], (g$r[ss])*100, lwd=1, col=col, lty='dotted')
    lines(g$t[ss], (g$r[ss])*200, lwd=1, col=col, lty='dashed')
    if(label){ # label last point in this one too?
      pp <- (g$r[n])*100
      dy <- ifelse(pp < 50, 0, 0) # hack to avoid label collision for small pp
      showxy(g$t[n],pp,dy=dy, show.point=FALSE)
    }
  }
}

# plot a few examples
growex2q <- function(yrs=700, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12, lag=FALSE){
  # Set up plot axes
  plot(NULL,type='n', xlim=c(0,1.2*yrs), ylim=c(0,100), xlab='T (years)', ylab='%',
       main='Population structure, assuming L is occasionally sexual')
  asex  <- grow1(yrs, a,         by, k, f, p0) # *a*sex simulation w/ default params
  sex00 <- grow2q(yrs, a, bx=0.0, by, k, f, p0) # sanity check: new w/ bx=0 should = old
  sex01 <- grow2q(yrs, a, bx=0.1, by, k, f, p0) # sim w/ more sexual offspring
  if(lag){
    sex02 <- grow2q(yrs, a, bx=0.1, by, k, f, p0, lag=T) # repeat w/ lag
    leg02 <- 'Same, lagged.'
    cat('Summary of r minus rlagged:\n')
    print(summary(sex01$r-sex02$r))
  } else {
    sex02 <- grow2q(yrs, a, bx=0.2, by, k, f, p0) # sim w/ even more
    leg02 <- '%L: bx=0.2, by=0.5'
  }
  #sex022<- grow2p(yrs, a, bx=0.22, by, k, f, p0) # sim w/ even more
  plot1(asex,  'yellow')
  plot2q(sex00, 'blue', allele=F)
  plot2q(sex01, 'darkgreen')
  plot2q(sex02, 'red', label=T)
  legend('bottomright',
         legend=c('%L: bx=0.0, by=0.5', '%L: bx=0.1, by=0.5', leg02, 'Allele freqs'),
```

```
        col=   c('blue',               'darkgreen',          'red', 'black'),
        lwd=c(2,2,2,1), lty=c(1,1,1,3), bty='n', cex=.9)
}
```

```
fpath <- '/Users/ruzzo/Documfpathents/g/projects/thaps/Thaps_7_strains/scripts/larrys/asex/figs'
```

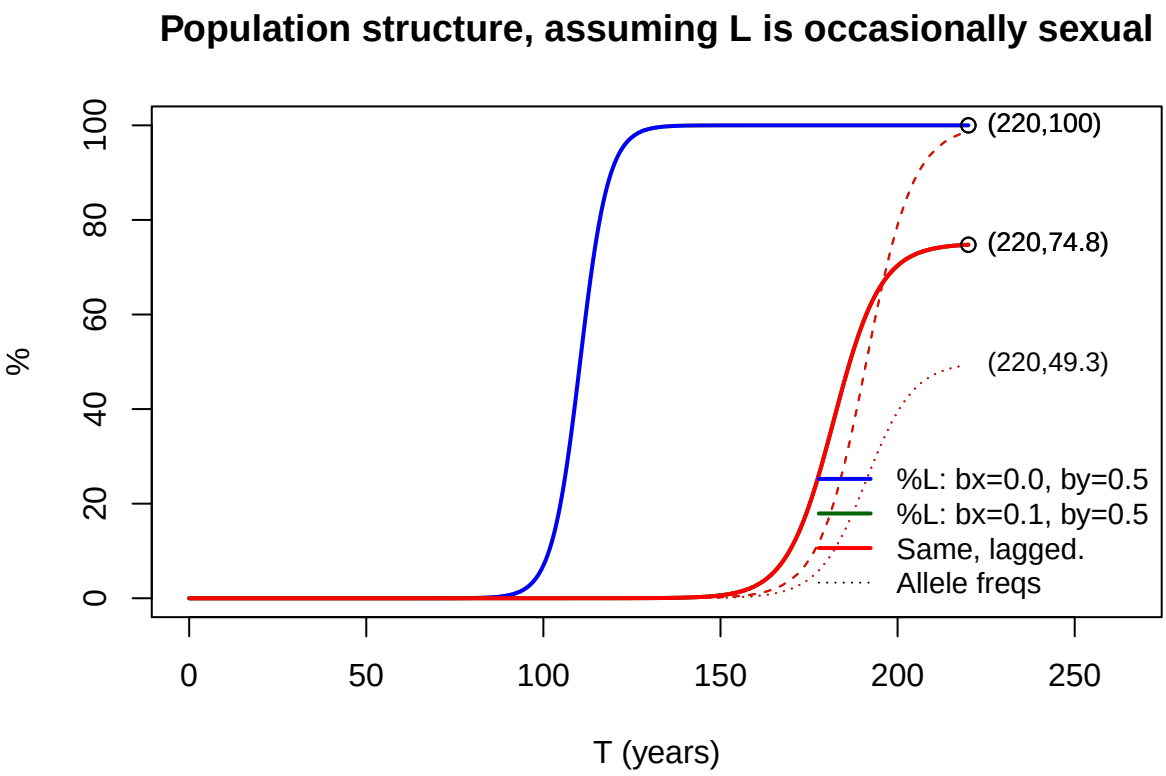
```
growex2q()
```



Reassuringly, this reproduces the earlier figure.

Low rate of sex introduces a lag, loosely modeled above. [Specifically, the genotype present in a cell at meiosis was fixed at that cell's *previous* meiosis/fertilization, not at the previous epoch; see comments in the code above for more details.] Does is change the results much?

```
growex2q(yrs=220, lag=T)
```




```
## Summary of r minus rlagged:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 0.000e+00 1.400e-08 1.327e-04 5.277e-05 1.024e-03
```

Seemingly, not much change; invisible on this scale.

This reproduces the key features of the earlier figure – assuming non-zero meiotic rate (non-blue curves), L rises to a stable fraction of the population $< 100\%$, and all heterozygous positions in L rise to 50% allele frequency within the non-L population. Not shown, but still true – homozygous positions in L rise to 100% in non-L. I.e., pre-existing non-L alleles are totally washed out of the population.

4. Model 4: Simple Mendelian Recessive

To recap: Under the simple assumptions modeled above, (Model 1) an obligate asexual genotype will sweep the world, whereas (Model 2) a facultative sexual genotype with reduced frequency of meiosis will sweep to a large fraction of the population, but stably coexist with its sexually derived offspring. Our observational data is plausibly consistent with a transient state[1] of (M1), and, at first blush, with (M2), although that doesn't hold up vs the more detailed model (M2prime). I claim that an additional *very plausible* genetic assumption is enough to render the "facultative sexual" model even less likely.

What might cause an increase in the average time between meiotic events? One model, perhaps the simplest, is that the increased interval is a simple Mendelian recessive trait—some single gene involved in meiosis has acquired a non-wildtype allele that, when homozygous, reduces the frequency of meiosis. I.e., assume that L is homozygous for the recessive allele of this gene, call it "ss" ("SlowSex"). ("Recessive" as opposed to dominant or additive is the more conservative choice I believe; either of these alternatives would sweep more rapidly, making the intermediate state less likely to be observed; considered briefly below as well.)

[1] Continued existence of the H isolates also may depend on factors completely outside the scope of the models considered here. For example, geographical isolation, esp. of the Italian isolate, may be at play, as may increased fitness to local environmental conditions.

As intuition for how this would affect thaps evolution, note that the L-genotype, sexual or not, has a growth advantage over both the wildtype and over heterozygotes. But if L remains sexual, it would continuously inject the ss allele into the non-L gene pool. (We have previously argued that all sexual offspring of L will have a recognizably non-L genotype.) As seen in the two models above, L rises exponentially in frequency. When its frequency is small, its contribution to the non-L gene pool is negligible, but after L attains a substantial representation, the proportion of ss injected into the non-L population becomes significant, after which *non-L homozygotes* will emerge, and they will have at least the same growth advantage as L has over wildtype. Thus, the ss allele will increase in frequency *among non-L genotypes*, eventually dominating the population. Furthermore, the L genotype is at a competitive disadvantage when compared to homozygous non-L, since sexual offspring of L genotypes are never of that genotype, whereas sexual offspring of homozygous non-L sometimes retain homozygosity. Said more simply, the "ss" allele, while recessive, is evolutionarily favored, at least in the short term, and will sweep to fixation, once given a kick-start by creation of the homozygous L-genotype. It's surprisingly easy.

The simulations below support this intuition, with the ss allele rapidly fixing in the population, and with the additional nuance that the L genotype is only ephemerally dominant in the population, and decreasingly so as the gap between b_x and b_y shrinks. Consequently, our observed data with 5 of 7 samples matching the L genotype seems more likely to reflect the "obligate asexual" model (see footnote 1 again) than the "facultative sexual, Mendelian recessive, sampled at just the right time" model. However, I admit that the data are not conclusive.

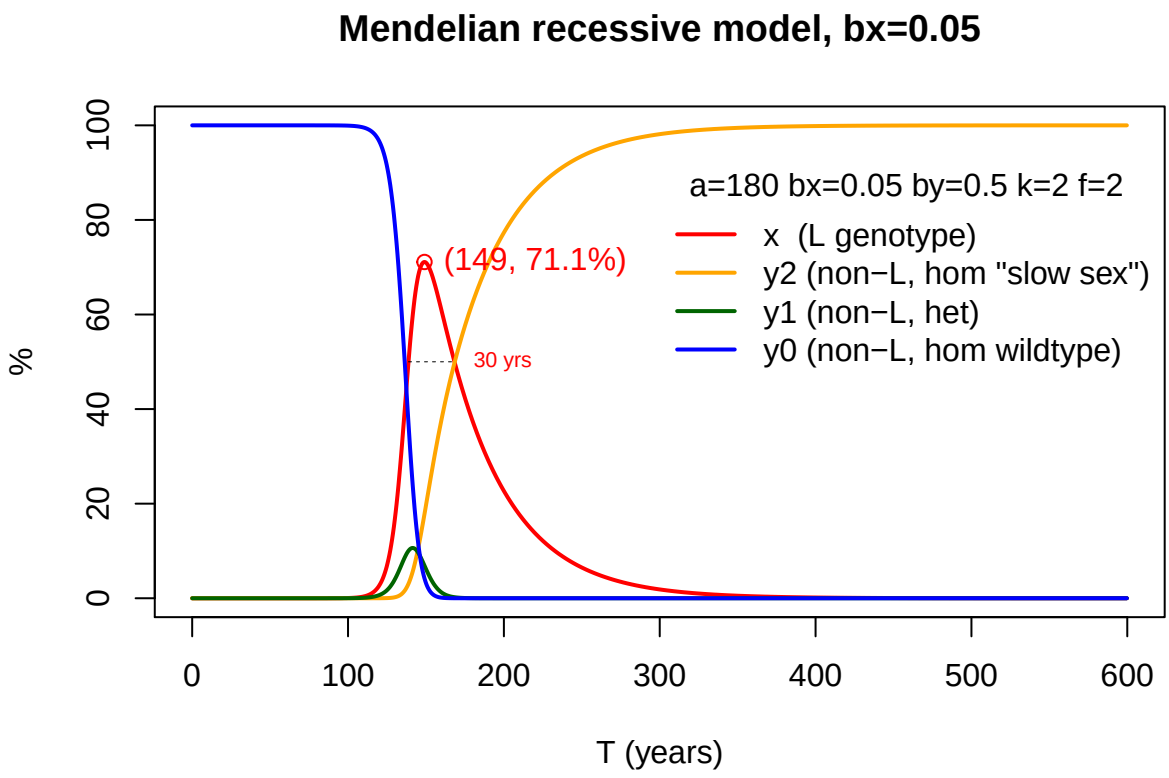
```
# grow4 simulates 'yrs' years of growth in the 4-component model where L's lower meiotic
# rate is assumed to be driven by presence of homozygous recessive form of some one gene.
# Gene flow from L to non-L is tracked, separately accounting for non-L types having 0, 1
# or 2 copies of the recessive allele. 0 and 1 copy cells are assumed to have b_y meioses
# per year; 2 copies (incl L) have b_x.  All have equal mitotic rates.
grow4 <- function(yrs, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12, recessive=TRUE){
  n <- round(yrs*a/k) # number of epochs to simulate
  x <- numeric(n)     # proportion in L
  y0<- numeric(n)     # proportion non-L with 0 copies
  y1<- numeric(n)     # proportion non-L 1 copy
  y2<- numeric(n)     # proportion non-L 2 copies
  betax <- bx/(a/k)   # meioses per epoch in L and y2
  betay <- by/(a/k)   # meioses per epoch in non-L y0, y1
  x[1]  <- p0         # initial proportion in L
  y2[1] <- 0
  y1[1] <- 0          # p0 is also reasonable here, but <- 0 is conservative
  y0[1] <- 1-x[1]
  for(i in 2:n){
    # total meiotic fraction
    z <- (x[i-1]+y2[i-1])*betax + (y1[i-1]+y0[i-1])*betay
    # fraction of recessive chromosomes among mieotics
    p <- (2*(x[i-1]+y2[i-1])*betax + 1*y1[i-1]*betay)/(2*z)
    q <- 1-p
    if(i<0){ # debug
      cat('yrs=',yrs,'a=',a,'k=',k,'n=',n,'i=',i,'betax=',betax,'betay=',betay,
          'z=',z,'p=',p, '1-q=',1-q,'x[i-1]=',x[i-1],'\n')
    }
    # Despite all the comments about "recessive", I also wanted to see what happens if the
    # trait is dominant, which just involves changing the "beta" coefficient in the y1
    # calcuation below.
    if(recessive){
      beta1 <- betay
    } else {
      beta1 <- betax
```

```
    }
    # new population fractions, assuming random mating
    x[i]  <-  x[i-1]*(1-betax)*2^k
    y2[i] <- y2[i-1]*(1-betax)*2^k +   p^2 * z * f # \
    y1[i] <- y1[i-1]*(1-beta1)*2^k + 2*p*q * z * f #  }-- Hardy Weinberg proportions
    y0[i] <- y0[i-1]*(1-betay)*2^k +   q^2 * z * f # /
    # re-normalize:
    w <- x[i]+y2[i]+y1[i]+y0[i]
    x[i]  <-  x[i]/w
    y2[i] <- y2[i]/w
    y1[i] <- y1[i]/w
   #y0[i] <- y0[i]/w
    y0[i] <- 1.0-x[i]-y2[i]-y1[i] # slightly more robust to roundoff errs, I think
  }
  t <- ((1:n)-1)/a*k # rescale epoch numbers to years
  return(list(x=x,y2=y2,y1=y1,y0=y0,t=t,yrs=yrs,bx=bx,by=by,a=a,k=k,f=f))
}

# plot4 plots a growth series g (as returned by grow4). Plots all 4 subpopulations.
plot4 <- function(g, main.sub.label='', where='topright', showLMax=TRUE, recessive=TRUE){
  # Set up plot axes
  main.label <- paste('Mendelian', ifelse(recessive, 'recessive', 'dominant'), 'model')
  if(main.sub.label != ''){
    main.label <- paste(main.label, main.sub.label, sep=', ')
  }
  plot(NULL,type='n', xlim=c(0,1.0*g$yrs), ylim=c(0,100), xlab='T (years)', ylab='%',
       main=main.label)
  n <- length(g$t)
  if(n < 2048){
    lines(g$t,  g$x*100, lwd=2, col='red')
    lines(g$t, g$y2*100, lwd=2, col='orange')
    lines(g$t, g$y1*100, lwd=2, col='darkgreen')
    lines(g$t, g$y0*100, lwd=2, col='blue')
    if(!recessive){
      lines(g$t,(g$y1+g$y2)*100, lwd=1, col='orange', lty='dashed')
    }
  } else {
    # For long series, sufficient, and much faster, to plot a subsequence
    k <- floor(n/1024)
    s <- seq(from=1,by=k,to=n)
    lines(g$t[s],  g$x[s]*100, lwd=2, col='red')
    lines(g$t[s], g$y2[s]*100, lwd=2, col='orange')
    lines(g$t[s], g$y1[s]*100, lwd=2, col='darkgreen')
    lines(g$t[s], g$y0[s]*100, lwd=2, col='blue')
    if(!recessive){
      lines(g$t[s],(g$y1[s]+g$y2[s])*100, lwd=1, col='orange', lty='dashed')
    }
  }
  # print((g$y1[n]+g$y2[n])*100)
  legend(where, lwd=2, bty='n', inset=c(0,0.1),
         title=paste('a=', g$a, ' bx=', g$bx,' by=', g$by, ' k=', g$k,' f=', g$f, sep=''),
         legend=c('x  (L genotype)',
                  'y2 (non-L, hom "slow sex")',
```

```
                  'y1 (non-L, het)',
                  'y0 (non-L, hom wildtype)'),
        col=   c('red', 'orange', 'darkgreen', 'blue'))
  if(showLMax){
    # annotate the max x point
    i <- which.max(g$x)
    points(g$t[i],g$x[i]*100,col='red')
    text(g$t[i], g$x[i]*100,
         paste('(', round(g$t[i]), ', ', round(g$x[i]*100,1),'%)', sep=''),
         pos=4, col='red')
    # and if x exceeds 50%, show where, & for how long
    maj <- which(g$x>=0.5)
    majn <- length(maj)
    if(majn > 0){
      majt <- g$t[maj[c(1,majn)]]         # 1st, last times
      lines(majt,c(50,50),lwd=0.5,lty=2) # draw a line there & say how many yrs
      text(majt[2], 50, paste(round(majt[2]-majt[1]),'yrs'),pos=4, cex=.7,col='red')
    }
  }
}
# plot an example
growex4 <- function(yrs=600, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12, where='topright',
                    verbose=FALSE, recessive=TRUE, main.sub.label=''){
  ex <- grow4(yrs, a, bx, by, k, f, p0, recessive)
  plot4(ex, where, recessive=recessive, main.sub.label=main.sub.label)
  if(verbose){
    n <-length(ex$x)
    print(rbind(
      'x-p0'=ex$x[c(1:5,n)]-p0,
      y2  =ex$y2[c(1:5,n)],
      y1  =ex$y1[c(1:5,n)],
      '1-y0'=1-ex$y0[c(1:5,n)]
    ))
    #print(1-ex$y0[1:5]-ex$x[1:5]-ex$y1[1:5]-ex$y2[1:5])
    #print(summary(1-ex$y0-ex$x-ex$y1-ex$y2))
  }
}
```

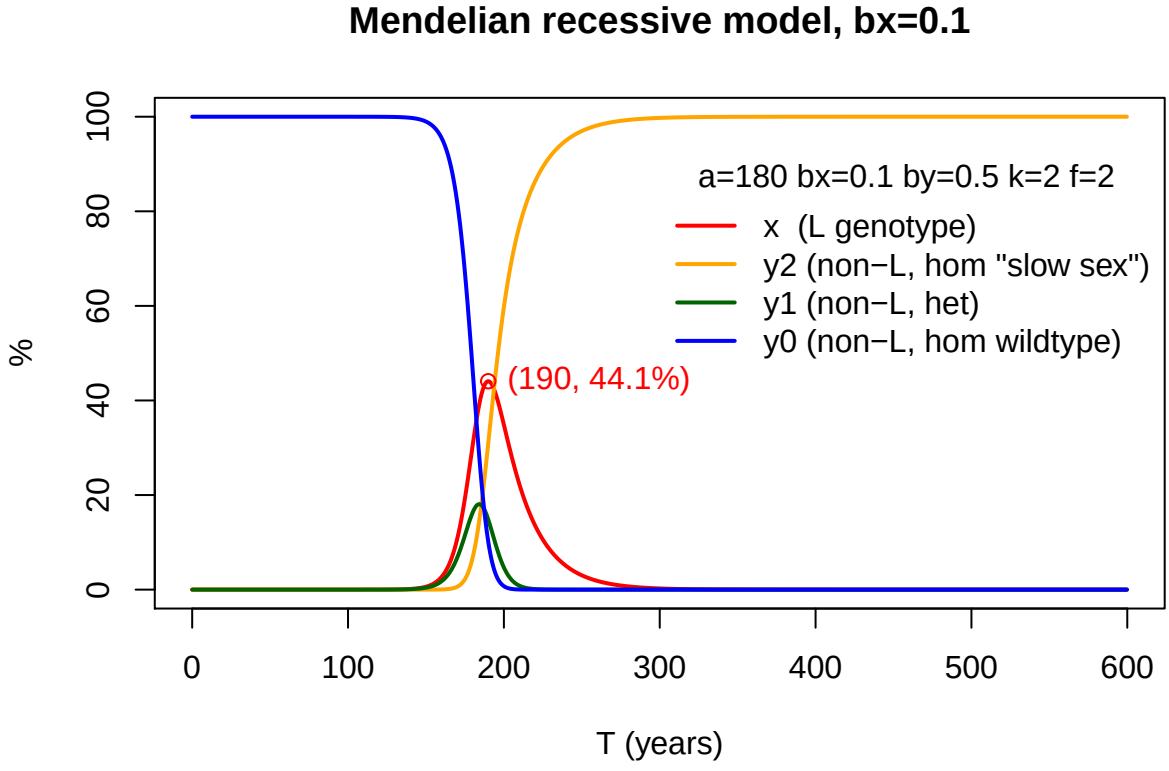
The examples shown below all use the same parameters $a = 180, b_y = 0.5, k = 2, f = 2, p_0 = 10^{-12}$ as in the earlier sections; they differ in that they model a hypothetical Mendelian recessive allele which, when homozygous, lowers the meiotic rate to $b_x = 0.05, \ldots, 0.24$, (vs $b_y = 0.5$ for heterozygotes and for homozygous wild type). In all, the wild-type remains dominant for a few centuries, then crashes rapidly as the L-genotype starts reaching significant levels, but the L-genotype's rise is quicky reversed as homozygous non-L genotypes out-compete and/or interbreed with it. The L-genotype exceeds 50% of the population only in the first simulation, and only briefly (≈ 30 years); note that this simulation assumes an average of 20 years between meioses, ten-fold less frequent than the default model of once every 2 years.

```
growex4(bx=0.05, main.sub.label='bx=0.05')
```

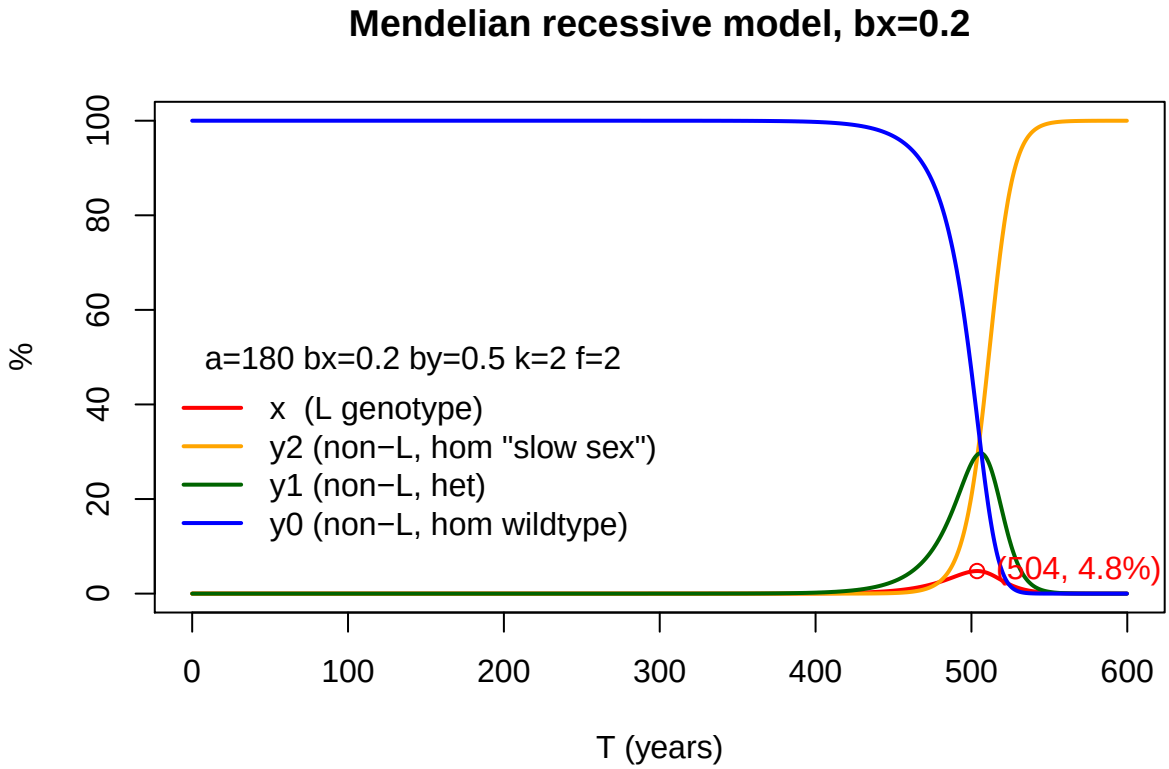


Raising b_x from 0.05 to 0.24 (the next three plots) delays the transition and lowers the extent of L expansion, but otherwise recapitulates the phenomena seen above.

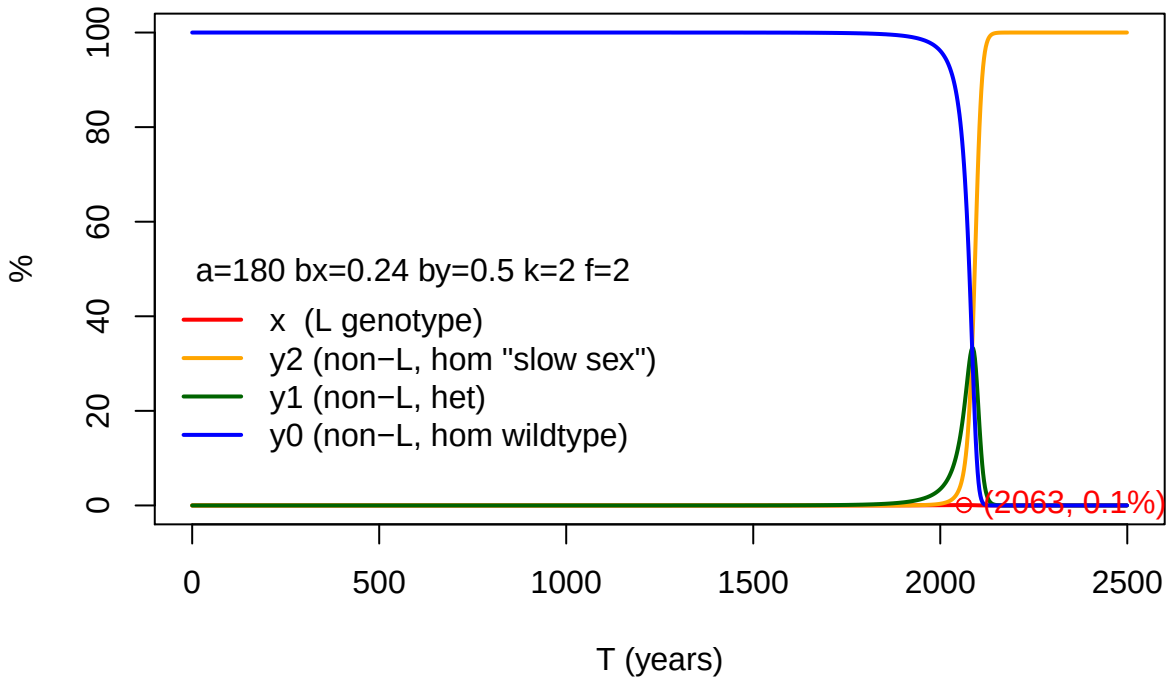
```
growex4(bx=0.1, main.sub.label='bx=0.1')
```



```
growex4(bx=0.2, main.sub.label='bx=0.2',where='bottomleft')
```



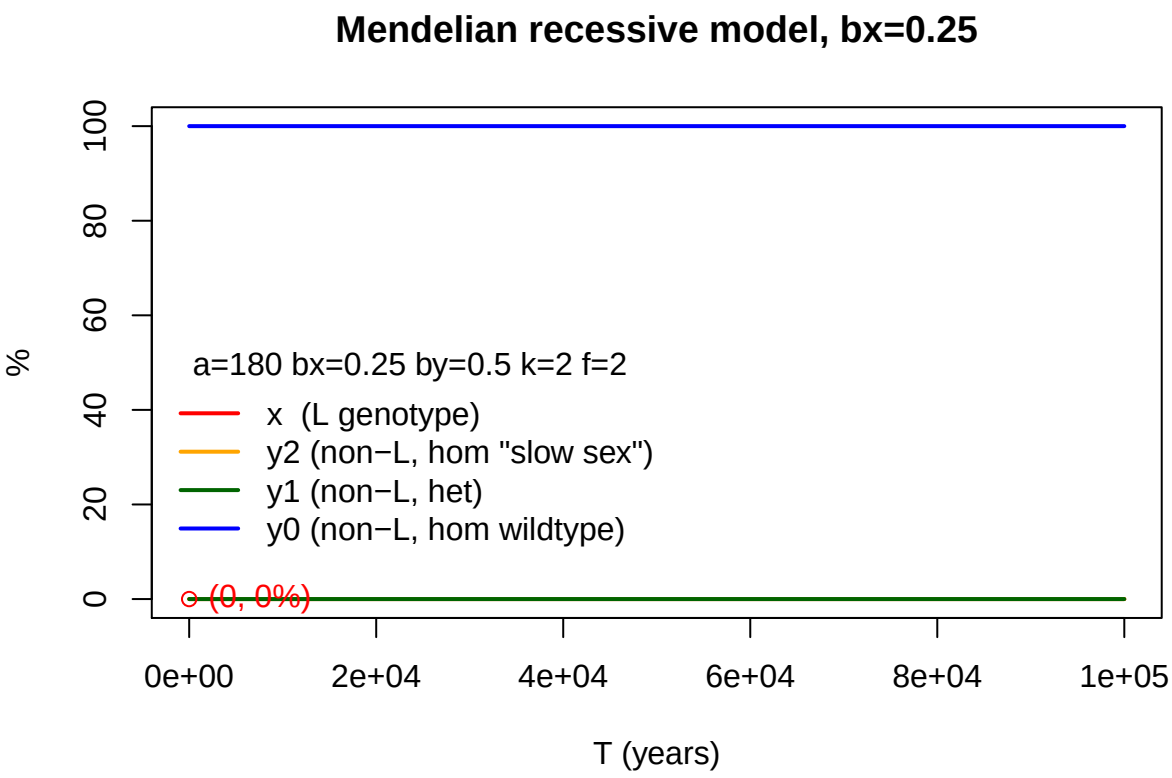

```
growex4(bx=0.24, main.sub.label='bx=0.24', where='bottomleft',yrs=2500)
```



There is, however, a qualitative change in character as b_x increases from 0.24 to 0.25. As in the examples

above, $b_x < b_y$ implies a growth advantage for the homozygous recessive genotypes since they undergo (slow) meiosis less frequently. However, for any $b_x > 0$, this growth advantage is at least partially offset by meiotic disruption of homozygosity. E.g., when $b_x = 0.25$, one quarter of the homozygous recessive population undergoes meiosis every year, mostly mating with cells *not* carrying the ss allele (until its allele frequency rises). At $b_x = 0.25$, these effects apparently balance, preventing the ss allele frequency from growing appreciably, even over a 100 thousand year span:

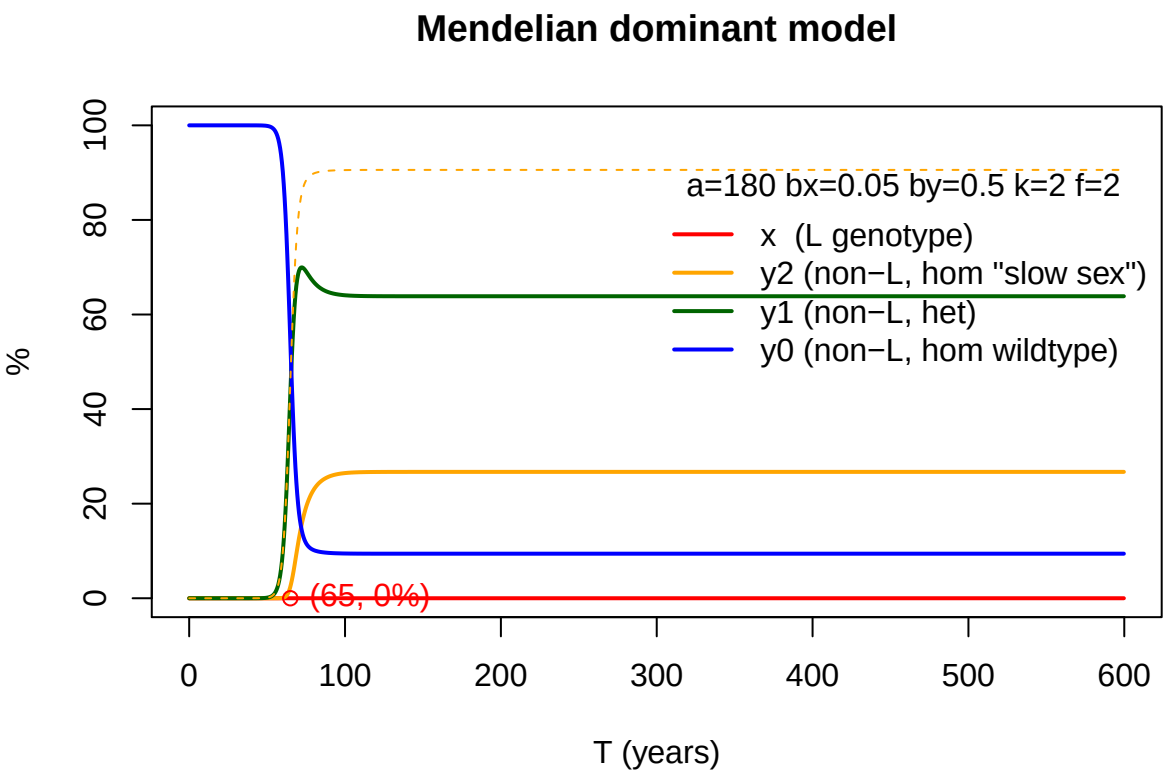
```
growex4(bx=0.25, main.sub.label='bx=0.25', where='bottomleft',yrs=100000,verbose = TRUE)
```




```
##              [,1]          [,2]          [,3]          [,4]          [,5]
## x-p0 0.000000e+00 -1.413639e-27 -2.827277e-27 -4.240916e-27 -5.654555e-27
## y2   0.000000e+00  6.963788e-28  1.396643e-27  2.100802e-27  2.808868e-27
## y1   0.000000e+00  2.785515e-15  5.571031e-15  8.356546e-15  1.114206e-14
## 1-y0 9.999779e-13  1.002753e-12  1.005529e-12  1.008305e-12  1.011080e-12
##               [,6]
## x-p0 -1.783971e-20
## y2    1.869332e-12
## y1    3.513800e-08
## 1-y0  3.514087e-08
```

If we instead assume that slow sex is a *dominant* trait then, in comparison to the first parameterization plotted in this section, dominance speeds the transition by about 50 years, and suppresses growth of the L genotype, but does not result in fixation of the ss allele, as seen in the plot below. Instead, the wild type persists at a low but stable level. Perhaps because heterozygotes get the growth advantage, hets persist in the population, and HWE then yields a few homozygous wild-type cells. [[Or maybe there's a bug in my code...]] (Out of laziness, the plot below is somewhat mislabeled, but the key new feature is the dashed orange line, which records the percentage of the population with 1 *or* 2 copies of the ss allele. Key point is that it does *not* reach 100%.)

```
growex4(bx=0.05,recessive=FALSE)
```



I expect an additive model (i.e., the heterozygote phenotype is intermediate between the homozygotic extremes) to be intermediate in effect – i.e., intermediate transition time and perhaps intermediate retention of the wildtype – but I have not checked this.

Hitchhiking

(The term "Hitchhiking" appears in the population genetics literature. As I understand it, the usage there is to describe the elevation of allele frequencies of neutral or perhaps even somewhat deleterious alleles that happen to be in linkage disequilibrium with a gene being positively selected, thus responsible, e.g. for LoH in a selective sweep. See Gillespie Section 4.2 (and the related notion of "genetic draft" in 4.3). The process described here is very similar, and so I'm going to borrow the name, but there is some risk of coinfusion since *un*linked loci are also involved, so I may need to rename this in the end. . .)

This section looks at population-wide allele frequencies. The hypothesis is that even under the Mendelian recessive model (where the "slow-sex" allele comes to dominate the population, but the L-genotype peaks, then fades), all other alleles in the L-genotype, even ones *not* in linkage disequilibrium with "slow-sex", will "hitchhike" on the success of slow-sex, reaching relatively high frequencies in the global population. This appears to be correct. Thus, in contrast to the neutralist expectation that most variation will be rare, all standing variation that happened to be present in L will be lifted to relatively high frequencies in the population (but in HWE, i.e., sometimes het and sometimes hom for either allele, as opposed to the uniformly het states seen in L). This is also in contrast to the usual . . .

This should be readily visible in a moderate size population sample – a hundred thousand SNPs at 20% (say) allele frequency. (Unfortunately, we have no such sample. . .)

On the other hand, although I did not attempt to model it, I expect LoH in the vicinity of the slow-sex allele – i.e., the hallmark of a classic selective sweep. (hmmm. is that right? if L is het in the vicinity, won't those 2 haplotypes *both* be amplified?) AND HOW IS THE HITCHHIKING EFFECT DIFFERENT FROM OTHER SELECTIVE SWEEPS??? [[I'm still mulling all this; recently-added section 5, is partial answer. "LoH near ss" - yes, in the sense that pre-existing non-L variation near ss will be reduced, but BOTH

neighboring haplotypes will be amplified, assuming they are different, which isn't quite like the classic sweep. I think hitchhiking ≠ sel sweep in that unlinked loci are involved, even other chromosomes.]]

```
# Like grow4 above, this version assumes slow sex is recessive.  Then tracks, within the
# non-L-population, the allele frequencies of (a) the recessive ss allele, and (b) an
# UNLINKED SNP present in L.  E.g., at a position that is polymorphic in L, does growth of
# L + spread of ss out of L also pull unlinked L-alleles out into the non-L population?
# (This counts "identity by descent", i.e. ancestry traces back to L; frequency will
# obviously be higher if it preexisted in non-L.)
#
# As above, the variables "x, y", below, record (over time) the proportion of the
# population in one of the following 2 genotype classes:
#
#   x: the L-genotype
#   y: all non-L (e.g., original wildtypes, L-inbreds and L-outcrosses)
#
# Note that in x, all positions are either homozygous for the same nucleotide across all
# cells, or heterozygous for the same pair of nucleotides across all cells.  In y, which
# is initially all but p0, we could see any allele frequencies.  A question of interest,
# in the case that x grows, is whether y can retain unique allele frequencies or whether
# mating of x with itself or with y will push y's allele frequency structure to become
# more L-like.  This simulation suggests that yes, this happens, but to a more limited
# extent that in the complex-trait model (grow2 etc.)
#
# Variable 'ss' below tracks freq of the ss allele in the y (non-L) pop.  Similarly, 'ul'
# tracks the freq (in y) of one allele at an unlinked SNP position.  Note that ul <= 0.5,
# since this is tracking only one of the 2 alleles at the SNP.  The other allele follows
# the same equations.  E.g., if ul rises to 0.4, then 40% of chromosomes in y will carry
# one of the 2 forms of the SNP, and 40% will carry the other, while whatever allele(s)
# were initially present in non-L will constitute the remaining 20% (in the same
# proportions as initially, if there were more than one).  Likewise, an unlinked
# homozygous position in L will rise to 2*ul frequency in non-L, regardless of which non-L
# allele(s) were initially present.
#
# As of 3/17 code review, comments above and in grow4q seem consistent with code.
#
# "complexity" param is a late addition; =2 means recessive alleles @ 2 loci
#
grow4q <- function(yrs, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12, complexity=1){
  n   <- round(yrs*a/k)# number of epochs to simulate
  x   <- numeric(n)    # proportion in L
  y   <- numeric(n)    # proportion non-L
  ss  <- numeric(n)    # allele freq of "slowsex" in non-L
  ul  <- numeric(n)    # allele freq of an unlinked "identical by descent L SNP in non-L
  betax <- bx/(a/k)    # meioses per epoch in L
  betay <- by/(a/k)    # meioses per epoch in non-L
  x[1]  <- p0          # initial proportion in L
  y[1]  <- 1-x[1]
  ss[1] <- 0
  ul[1] <- 0
  for(i in 2:n){
    # Tracking L admixing into y's allele frequency structure: Consider one epoch. Picture
    # x,y as cell *counts* in the population (in units of the total pop size to be
    # precise). Count chromosomes (2 per cell, of course) being added to y's gene pool as
```

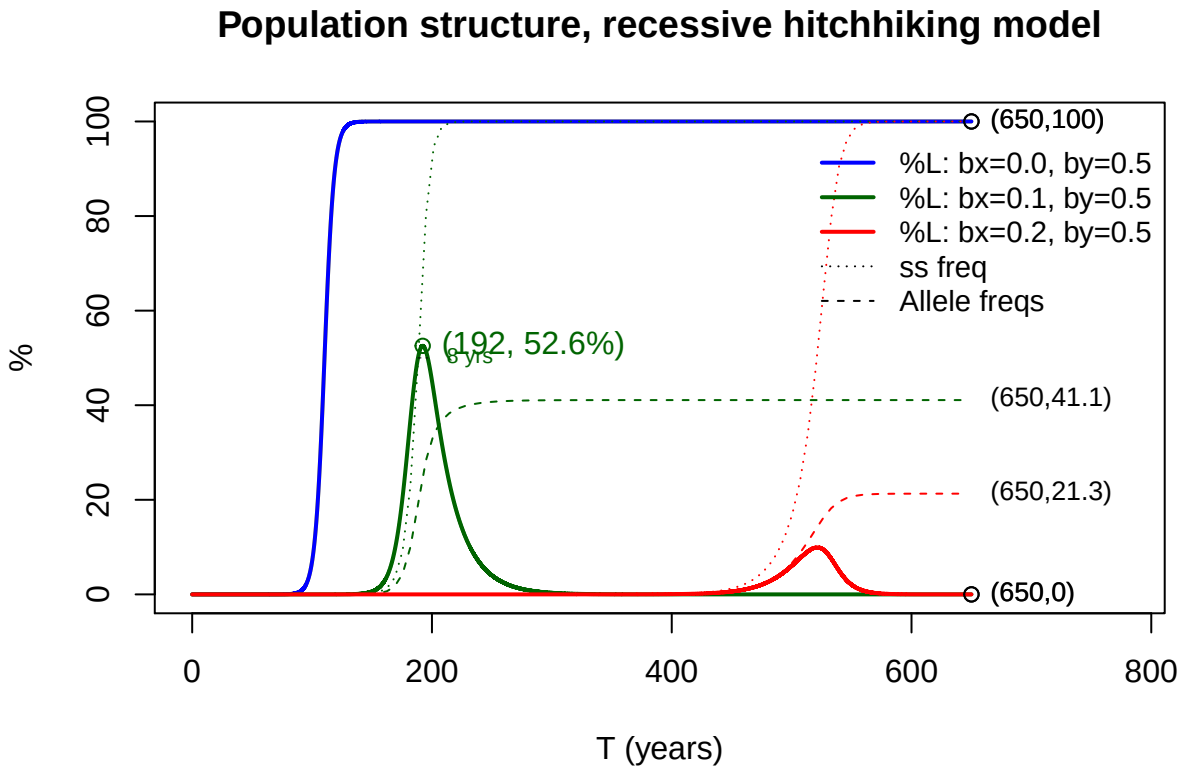
```
  # a result of both mitosis and meiosis of y cells, and meiosis of x cells. Mitosis
  # within x contributes to x, not y.  Perhaps conceptually clearer to say "count
  # physical copies of one haploid genome, of which there are 2 per cell".  I've tried
  # to clean up language to reflect this view but there may be some lingering comments
  # where "chromosome" should be replaced by haplocopy
  #
  # the hc.-.-- variables below are all triples of haplocounts at the end of the i-th
  # epoch.  E.g., hc.y.meio gives haplocounts arising from meioses in y. The first
  # component gives the total haplocount with that origin.  Second component is the
  # expected number carrying the 'tracked' variant (unlinked to ss), and the 3rd
  # component is the expected number of them carrying the ss allele.  Expected values
  # all assume HWE based on frequencies at the end of the previous epoch.
  #
  # E.g., of the x[i-1] L cells, a 1-betax fraction are exclusively mitotic during the
  # epoch, giving rise to 2^k offspring, each having 2 haplocopies, both with the ss
  # allele and one with the tracked variant, so:

  hc.x.mito <- x[i-1] * (1-betax) * 2^k * c(2, 1, 2)

  # The remaining betax fraction are meiotic.  I found this case to be somewhat
  # confusing at first; I think the following view is the simplest way to approach it.
  # On average, each meiotic cell generates 2*f successfully fertilized/fertilizing
  # haploid gametes (of course, organized into f diploid cells, consistent with our
  # assumption that there are f meiotic offspring per meiotic cell on average).  Each
  # gamete carries one haplocopy, all with the ss allele and half with the tracked
  # variant, so:

  hc.x.meio <- x[i-1] *    betax  * 2*f * c(1, 0.5, 1)

  # Counts for haplocopies arising from mitoses/meioses of non-L cells (y) follow
  # similar logic, with two key differences.  First, these cells possess the ss allele
  # in HWE proportions, based on 'ss' from the end of the previous epoch.  Second, the
  # ss[i-1]^2 fraction that are homozygous for the recessive slow sex allele undergo
  # meiosis at the slower betax rate; the rest at betay.  (NB: pushing 2 from 2*f into
  # the c(,,) triple increases code similarity between mito/meio cases.) So:

  tt <- c((1-ss[i-1])^2, 2*ss[i-1]*(1-ss[i-1]), ss[i-1]^2)
  if(complexity == 1){
    hc.y.mito <- y[i-1] * 2^k * ((1-betay) * c(2, 2*ul[i-1], 0) * tt[1] +
                                 (1-betay) * c(2, 2*ul[i-1], 1) * tt[2] +
                                 (1-betax) * c(2, 2*ul[i-1], 2) * tt[3]
                                )

    hc.y.meio <- y[i-1] *   f * (    betay * c(2, 2*ul[i-1], 0) * tt[1] +
                                     betay * c(2, 2*ul[i-1], 1) * tt[2] +
                                     betax * c(2, 2*ul[i-1], 2) * tt[3]
                                )
  } else {
    if(complexity > 2){ stop('unsupported complexity') }
    ## COMPLEXITY 2 CASE NEEDS CAREFUL VETTING
    # modeling allele counts at 2 loci jointly, i.e. 0:2 x 0:2.  uu[.,.] has all 9
    # probabilities, but because, e.g., the 0,1 and 1,0 cases have equal probability,
    # both loci will have the same dynamics, so suffices to summarize one of them.
```

```
      # vv[.] has the 5 anti-diagonal sums from uu[]; e.g. vv[i] is sum of cases where
      # there are i-1 allele copies in total at the 2 loci.  Again by symmetry, this is
      # equiv to (i-1)/2 copies at each.
      uu <- outer(tt, tt, FUN="*")  # 3 x 3 outer product of tt with self
      vv <- numeric(2*complexity+1)

      vv[1] <- uu[1,1]
      vv[2] <- uu[2,1]+uu[1,2]
      vv[3] <- uu[3,1]+uu[2,2]+uu[1,3]
      vv[4] <- uu[3,2]+uu[2,3]
      vv[5] <- uu[3,3]
      if(i==-99){ print(tt); print(uu); print(vv) }
      hc.y.mito <- y[i-1] * 2^k * ((1-betay) * c(2, 2*ul[i-1], 0/2) * vv[1] +
                                   (1-betay) * c(2, 2*ul[i-1], 1/2) * vv[2] +
                                   (1-betay) * c(2, 2*ul[i-1], 2/2) * vv[3] +
                                   (1-betay) * c(2, 2*ul[i-1], 3/2) * vv[4] +
                                   (1-betax) * c(2, 2*ul[i-1], 4/2) * vv[5]
                                  )

      hc.y.meio <- y[i-1] *   f * (    betay * c(2, 2*ul[i-1], 0/2) * vv[1] +
                                       betay * c(2, 2*ul[i-1], 1/2) * vv[2] +
                                       betay * c(2, 2*ul[i-1], 2/2) * vv[3] +
                                       betay * c(2, 2*ul[i-1], 3/2) * vv[4] +
                                       betax * c(2, 2*ul[i-1], 4/2) * vv[5]
                                  )
    }

    hc.newy <- hc.x.meio + hc.y.mito + hc.y.meio

    # The proportion of x in the population at the start of the next epoch is just the
    # number of x-type chromosomes produced by x-mitosis, normalized by the total number
    # of chromosomes:

    x[i] <- hc.x.mito[1]/(hc.x.mito[1] + hc.newy[1])
    y[i] <- 1 - x[i]

    # Sumilarly, ss and ul are the corresponding fractions of the new y population.

    ul[i] <- hc.newy[2]/hc.newy[1]
    ss[i] <- hc.newy[3]/hc.newy[1]
  }
  t <- ((1:n)-1)/a*k  # rescale epoch numbers to years
  return(list(x=x, y=y, ss=ss, ul=ul, t=t))
}
#zzz<-grow4q(1,complexity=2)
```

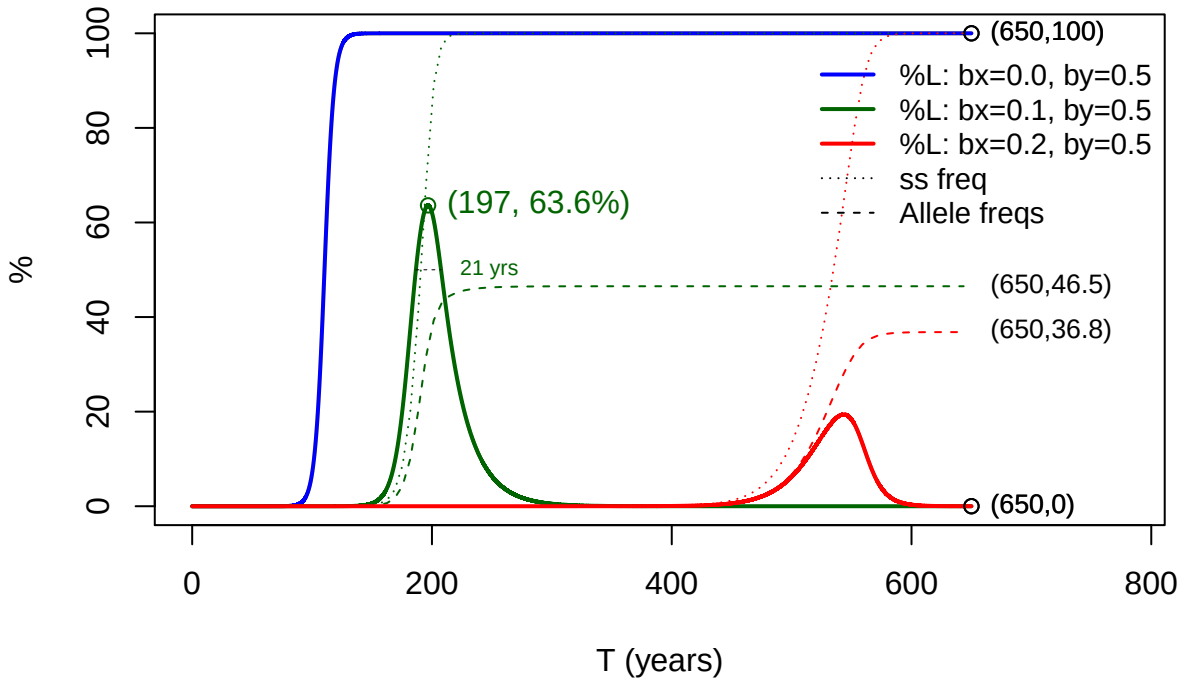
```
# plot4q plots a growth series g (as returned by grow4q) in specified color, line type
plot4q <- function(g, col, lty=1, label=F, showLMax=FALSE){
  n <- length(g$x)
  p <- g$x * 100                          # rescale x to percent
  lines(g$t, p, lwd=2, col=col, lty=lty)  # plot x vs t
  showxy(g$t[n], p[n])                    # annotate last point in series
  # now plot ss & ul as dotted, dashed, resp. (plot subseq of about 100 points for these,
  # otherwise dash/dot is obscured)
```

```
  sq <- seq(from=1, to=n, by=ceiling(n/100))
  lines(g$t[sq], (g$ss[sq])*100, lwd=1, col=col, lty='dotted')
  lines(g$t[sq], (g$ul[sq])*100, lwd=1, col=col, lty='dashed')
  if(label){ # label last point in this one too?
    pp <- (g$ul[n])*100
    dy <- ifelse(pp < 50, 0, 0) # hack to avoid label collision for small pp
    showxy(g$t[n],pp,dy=dy, show.point=FALSE)
  }
  if(showLMax){
    # annotate the max x point
    i <- which.max(g$x)
    points(g$t[i],g$x[i]*100,col=col)
    text(g$t[i], g$x[i]*100,
         paste('(', round(g$t[i]), ', ', round(g$x[i]*100,1),'%)', sep=''),
         pos=4, col=col)
    # and if x exceeds 50%, show where, & for how long
    maj <- which(g$x>=0.5)
    majn <- length(maj)
    if(majn > 0){
      majt <- g$t[maj[c(1,majn)]]        # 1st, last times
      lines(majt,c(50,50),lwd=0.5,lty=2) # draw a line there & say how many yrs
      text(majt[2], 50, paste(round(majt[2]-majt[1]),'yrs'),pos=4, cex=.7,col=col)
    }
  }
}

# plot a few examples
growex4q <- function(yrs=600, a=180, bx=0.2, by=0.5, k=2, f=2, p0=1e-12, complexity=1){
  # Set up plot axes
  if(complexity == 1){
    x.factor <- ''
  } else {
    x.factor <- paste(complexity, 'x ', sep='')
  }
  main.txt <- paste('Population structure, ', x.factor, 'recessive hitchhiking model', sep='')
  plot(NULL,type='n', xlim=c(0,1.2*yrs), ylim=c(0,100), xlab='T (years)', ylab='%',
       main=main.txt)
  cp <- complexity
  asex  <- grow1(yrs, a,         by, k, f, p0) # *a*sex simulation w/ default params
  sex00 <- grow4q(yrs, a, bx=0.0, by, k, f, p0, cp) # sanity check: new w/ bx=0 should = old
  sex01 <- grow4q(yrs, a, bx=0.1, by, k, f, p0, cp) # sim w/ more sexual offspring
  sex02 <- grow4q(yrs, a, bx=0.2, by, k, f, p0, cp) # sim w/ even more
  leg00 <- '%L: bx=0.0, by=0.5'
  leg01 <- '%L: bx=0.1, by=0.5'
  leg02 <- '%L: bx=0.2, by=0.5'
  #sex022<- grow2p(yrs, a, bx=0.22, by, k, f, p0) # sim w/ even more
  plot1(asex,  'yellow')
  plot4q(sex00, 'blue')
  plot4q(sex01, 'darkgreen', label=T, showLMax = TRUE)
  plot4q(sex02, 'red', label=T)
  legend('topright', inset=c(0,.05),
         legend=c(leg00,  leg01,       leg02, 'ss freq', 'Allele freqs'),
         col=   c('blue', 'darkgreen', 'red', 'black',   'black'),
```

```
        lwd=c(2,2,2,1,1), lty=c(1,1,1,3,2), bty='n', cex=.9)
}
growex4q(650)
```




```
growex4q(650,complexity=2)
```



5. Digression: Dominant/Recessive/Sweep/Obligate/Facultative

I think population genetics of advantageous recessive alleles in facultative sexuals is probably not a widely studied scenario. Here I want to write down some general comments just to clear my head and hopefully not overlook anything.

Case 1: advantageous dominant in an obligate sexual population. This is probably the simplest and most widely applicable case. The allele arises once. Initially it seems plausible that drift is a more significant force than selection, but once it has reached a level where drift-to-extinction is improbable, then expansion to fixation becomes likely, due to selective advantage. Genetic hallmark would be the classic LoH due to selective sweep: the haplotype in linkage disequilibrium (LD) with the trait would be carried to fixation with it. The LoH region will be larger or smaller depending on strength of selection, effective population size/number of generations to fixation, average inter-crossover distance, etc.

Case 2: ditto, but facultative sexual. This seems nearly identical to case 1. Calculating the size of LoH regions might need to account for mitoses per meiosis on a typical lineage since selection is always acting, but crossover-erosion of the LD blocks happens only at meiosis, of course.

Case 3: advantageous recessive in an obligate sexual population. Again, assume the allele arises once. Again drift is initially important, and important for much longer since by assumption this allele is neutral except in homozygotes. And, in a randomly mating population, homozygotes remain rare until the allele frequency in the population rises to a substantial fraction. Here's a seat-of-the-pants analysis. Neutral theory says only 1 in 2N neutral mutations ever fix, where N is the effective population size. So 1 in N reach 50% allele frequency, at which point 1/2 fix and 1/2 go extinct. So only something like 1 in N (or, being generous: 10 in N) advantageous recessives ever rise to a level at which selection begins to favor them.

Partial isolation of a small founder population with concomitant inbreeding might accelerate that, but only if the selective advantage of the homozygous recessive state outweighs the other deleterious consequences of inbreeding. In short, compared to cases 1 and 2, and assuming a similar strength of selection on the advantageous state, it seems like this case would both greatly reduce the chance of success and greatly increase the time to fixation (and consequently greatly reduce the size of the surrounding LoH region, so no "selective sweep" hallmark may remain).

Case 4: advantageous recessive in a facultative sexual population. (Perhaps I should say "unlimitedly facultative", i.e., can divide mitotically an unlimited number of times, unlike diatoms with shrinking frustule sizes...) This is the thaps case. It seems pretty different, since creation of one homozygote, say by inbreeding, creates a scenario like cases 1,2 where selection acts immediately. It will also have very different genetic hallmarks, as we've seen in earlier sections - either a clonal sweep, if the mutant is an obligate asexual, or strong penetration of even unlinked haplotypes into the full population, if the mutant retains occasional sex.

6. Summary

If the L-genotype retains facultative sex, our observation of it at 5 of 7 loci requires an implausible concatenation of circumstances, genetics, timing, and sampling. It seems quite a bit more plausible that it is an obligate asexual that has risen to dominance globally (with the H-isolates persisting by some characteristic outside of our model, such as biogeographic isolation, local adaptation and/or other selective advantage).

---

7. Overlap

Digression: is a Venn diagram of SNP overlaps useful? NY v IT v Wales.

```
#install.packages('VennDiagram')
library(VennDiagram)
```

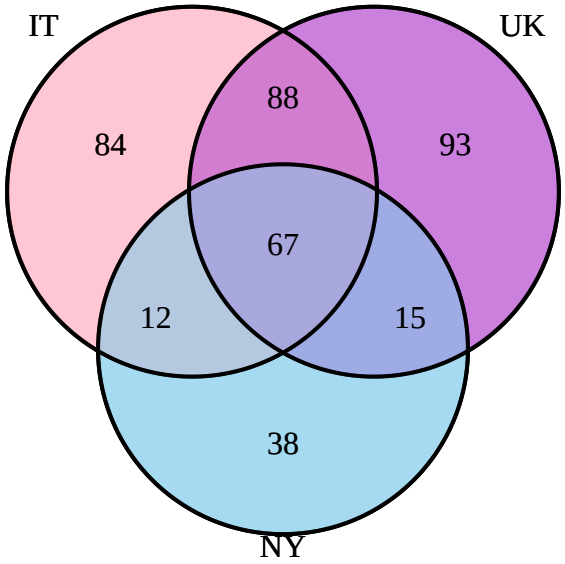


Figure 1: Venn Diagram of NY/IT/UK shared SNPs (in thousands)

```
n1   <- 38        # ni = area private to i; (1,2,3)=(NY,IT,UK)
n2   <- 84
n3   <- 93
n123 <- 67        # triple intersection
n12  <- 12 + n123 # dual plus triple intersect
n13  <- 15 + n123
n23  <- 88 + n123
a1 <- n1 + n12 + n13 - n123 # total area
a2 <- n2 + n12 + n23 - n123
a3 <- n3 + n13 + n23 - n123
```

```
pdf(file='figs/venn.pdf', width=3, height=3)
grid.draw(
  draw.triple.venn(area1 = a1, area2 = a2, area3 = a3, n12 = n12, n23 = n23, n13 = n13,
    n123 = n123, category = c('NY','IT','UK'),  rotation=2,
    # euler.d = TRUE, scaled = TRUE, ## seemingly do nothing
    fill = c("skyblue", "pink1", "mediumorchid")))
no.chatter <- dev.off()
```

Old and (mostly) in the way

The following discussion is largely subsumed by the foregoing; keeping it for now in case we revert. “Synopsis bullets 1 & 2” are still relevant; #3 is flawed in that it ignored the growth differential highlighted above (but suggests how to frame the models as differential equations instead of the discrete time simulations above).

Origin

This attempts to summarize summarize our thinking on obligate asexuality, starting from some recent emails, but also including much older material, (a) to put it in one place, and (b) letting Rstudio/RMarkdown/Pandoc convert the formulas to .docx format so I can paste (some of) it into the supplement.

Synopsis

1) It doesn't take multiple generations of sexual mixing to erase the L-clade's genotype; sexual offspring of a *single* cross will be recognizably distinct.
2) This process is essentially irreversible.
3) Hence, under the neutral assumption that L-clade and its hypothetical sexual descendants are equally fit, even infrequent sexual reproduction will result in an exponential decline in the proportion of L-clade cells in the population.

This seems inconsistent with global isolation of the L genotype, making "obligately asexual" the more parsimonious conclusion.

Supporting Arguments

Point (1). Sexual reproduction does not "slowly erode" the genomic traits we have used to define the L-clade (specific, large "SNP deserts"; shared SNPs; and near-absence of homozygous non-reference positions)—on the contrary, they will not survive even a single sexual generation. For example, both H-isolates have many thousands of SNPs that are not shared with the L-isolates, and offspring of a hypothetical L x H outcross would inherit many of these, since there is approximately a 50% chance of inheriting the novel non-reference nucleotide at any such position. Additionally, hypothetical offspring would be expected to be homozygous at approximately 50% of the thousands of positions at which L- and H-isolates share SNPs, and 50% of these would be the non-reference allele. Likewise, L-clade's large deserts aren't shared with the H-isolates, meaning that the later possess at least one non-reference haplotype across each of those regions, hence the large deserts likely would be fractured or absent in offspring. A hypothetical cross between two L-genotype cells *would* preserve the large deserts, since all 4 available alleles in the 2 parents are nearly identical, but *additional* loss of heterozygosity, including the creation of large tracts with many homozygous non-reference nucleotides, would inevitably occur: in each segment that is not an L-genotype desert, each parent has 2 alleles, the SAME 2 alleles, so offspring have a 50:50 chance of being identical to L or homozygous for one or the other of the parental alleles (hence, a desert). Further crosses / recombination / random assortment will break up these LoH/L-genotype blocks, erasing the block structure in the limit, but will still leave huge numbers of homozygous non-ref positions, as well as increasing the recombination density. In short, we easily have the power to discriminate between the L-genotype and a hypothetical cross of it with either an L- or a non-L genotype, again after as few as *one* sexual generation.

Point (2). Even if meiotic recombination were limited to one crossover per chromosome, the chance of preserving or reconstructing the L-genotype during sexual reproduction between L-genotype cells and/or their sexual descendants is astronomically small.

Point (3). Consider a simple model initially consisting entirely of L-genotype individuals that reproduce sexually at some rate (amidst the usual asexual proliferation). Given the above points, sexually derived offspring will *not* retain the L-genotype. For simplicity, we do not consider presence of non-L-genotypes other than those arising in this way, and we assume all genotypes are equally fit (but see below).

Let $x(t)$ and $y(t)$ measure the relative population sizes of L- and non-L-genotype (sexual descendants of L-genotype), resp., at time t. Assume initially $x(0) = 1, y(0) = 0$. Assume that the average interval between meioses along any cell lineage is s time units, and that for every cell entering meiosis, an average of f fertilized eggs are produced and survive. Under these assumptions, in a short time interval dt, the L-clade population size will become:

$$x(t+dt) = \left(1 - \frac{dt}{s}\right) x(t)$$

(1)

as a small fraction dt/s of the population undergoes meiosis, thereby ceasing to be recognizably L-clade (see point (1)), while

$$y(t+dt) = \left(1 - \frac{dt}{s}\right) y(t) + f \cdot (x(t) + y(t)) \cdot \frac{dt}{s}$$

(2)

as a similar fraction of the non-L population undergoes meiosis, and a fraction f of the meiotic cells (from both) are successfully fertilized (and again no longer L-clade).

Note that, assuming $s < \infty, x(t)$ is a decreasing function of t, and, depending on f, both $y(t)$ and the sum $x(t) + y(t)$ may grow or shrink. We assume that mitotic growth and/or predation/death maintain the total population at a fixed level (or follow typical bloom/bust cycles), but, by our neutral assumption, this affects x and y equally, so there is no need to explicitly model this. Instead, our goal is to find

$$P(t) = \frac{x(t)}{x(t) + y(t)},$$

(3)

the fraction of L-clade in the population as a function of time.

Based on (1), we have

$$x'(t) = \frac{d}{dt}x(t) = \frac{x(t+dt) - x(t)}{dt} = -\frac{x(t)}{s},$$

a differential equation whose solution is

$$x(t) = x(0)e^{-t/s} = e^{-t/s}.$$

Similarly, based on (2),

$$y'(t) = \frac{d}{dt}y(t) = -\frac{y(t)}{s} + \frac{f \cdot (x(t) + y(t))}{s}$$

which can be solved as follows:

$$y'(t) + \frac{1-f}{s}y(t) = \frac{f}{s}x(t) = \frac{f}{s}e^{-t/s}$$

Multiplying by $h(t) = e^{(1-f)t/s}$ and noting that $\frac{d}{dt}h(t) = \frac{1-f}{s}h(t)$:

$$h(t)y'(t) + \frac{1-f}{s}h(t)y(t) = \frac{f}{s}h(t)e^{-t/s}$$

$$h(t)y'(t) + h'(t)y(t) = \frac{f}{s}h(t)e^{-t/s}$$

$$(h(t) \cdot y(t))' = \frac{f}{s} e^{(1-f)t/s} e^{-t/s} = \frac{f}{s} e^{-ft/s}$$

$$h(t)y(t) = \int \frac{f}{s} e^{-ft/s} dt = -e^{-ft/s} + c$$

and based on the initial value $y(0) = 0$, we have $c = 1$, so

$$y(t) = (1 - e^{-ft/s})/h(t)$$

$$= (1 - e^{-ft/s}) e^{-(1-f)t/s}$$

$$= e^{-(1-f)t/s} - e^{-t/s}.$$

Substituting into (3):

$$P(t) = \frac{e^{-t/s}}{e^{-t/s} + e^{-(1-f)t/s} - e^{-t/s}}$$

$$= \frac{e^{-t/s}}{e^{-(1-f)t/s}}$$

$$= e^{-ft/s}.$$

I.e., the fraction of L-clade in the population declines exponentially with time.

Inserting plausibly conservative numbers for the model parameters, e.g., $f = 20\%$, and $t = 200$ years, we see that even if sexual reproduction were as rare as once every $s = 10$ years, then only $e^{-2*200/10} = e^{-4} < 2\%$ of the present-day population would retain the L-genotype.

[[ARE THESE REASONABLE NUMBERS FOR AN EXAMPLE??]]

However, the L-genotype has been the *T. pseudonana* genotype isolated in 5 of 7 tries, suggesting that it is a majority or prominent minority at ≥ 5 of 7 locations. To reconcile this with the above picture of an exponentially decaying L population, it could be that:

1) L's age (in comparison to f and s) is so small that the exponential decline is still modest; i.e., $t \ll 200$, $s \gg 10$, and/or $f \ll 20\%$.
2) While our model assumed *no fitness differences between L and its offspring* (more precisely, we assume that at most a fraction f of such offspring arise), perhaps L is such a precariously balanced genotype that it out-competes *all* of its sexual offspring ($f = 0$). A variant of this model is that L is a small minority everywhere, but is extremely adept at being brought into culture, and much more so than any of its sexual offspring.
3) L is an obligate asexual.

We feel that the example parameters are reasonably conservative, making scenario (1) unlikely. Also note that if the L-genotype were facultatively sexual, it would have been spawning non-L offspring at some rate since its emergence (a feature we did not attempt to model), making it even harder to establish the global presence we now see, unless some effect such as (2) is also at work.

We hypothesize that the L-genotype includes a particular combination of alleles that make it globally fit. This is, of course, an improbable occurrence, but as nature explores trillions of genotypes, such fortuitous

combinations will arise. Creating such a genotype that is in addition so much more fit than any of its sexual offspring, scenario (2), seems much less probable. Specifically, sexual offspring of a hypothetical L-L cross will differ from their parent by acquiring homozygosity in some regions. On average, this will retain beneficial and deleterious alleles equally, although overdominant loci suffer in either case. Is this enough to necessitate "inbreeding depression" in an otherwise globally fit genotype?

Hence, scenario (3) seems by far the most parsimonious explanation. L's global fitness is surprising, but that it should be so much more fit than any of its sexual offspring (scenario (2)) seems much less probable. On the other hand, rendering a fit genotype asexual is essentially trivial — just break one of the many key genes involved in meiosis or gamete recognition.

Notes from email

(These notes are essentially subsumed by the analysis above; included for completeness.)

9/8/17, 1st message:

I've been thinking more about it, so let me amplify. I said L x L yields the same genotype, but that is WRONG. The DESERTS will be preserved, since all 4 available alleles in the 2 parents are identical. But LARGE, NEW deserts will be created: in each segment that is not an L-clade desert, each parent has 2 alleles, the SAME 2 alleles, so offspring have 50:50 chance of being identical to L or homozygous for one or the other of the parental alleles (hence, a desert). Further crosses / recombination / random assortment will break up the sizes of these LoH blocks, erasing the block structure in the limit, but will still leave huge numbers of homozygous non-ref positions, as well as increasing the recombination density. In short, we easily have the power to discriminate between L-clade and a hypothetical cross of L with either an L or a non-L genotype, again after only *one* generation (even in the face of many homozygous nonref positions possibly being erroneously called SNPs).

The other thing to note about sex with L is that it's a one-way street: L cross anything yields non-L, while essentially no cross yields L. So, intuitively, any L-containing population in which L has sex, however rarely, will show an exponential decline (with time) in the proportion of L cells, since sex bleeds away L genotypes.

E.g., consider the simple scenario starting with a purely L population, say t=200 years ago, and assume L reproduces sexually once ever s=10 years, or, equivalently, 10% (1/s) of cells undergo meiosis every year, and that, say, the number of fertilized eggs that result is f=20% of the number undergoing meiosis. Picturing this as a synchronous process, when sex happens, the population temporarily shrinks (or grows, if f>1) from some number N of cells to .9N + (.1)*(.2)N = .92N, then grows back to the steady state size of N, BUT it is no longer purely L, it now is 90/92 = 97.7% L and 2.3% non-L. The same happens in the next year: 10% of the extant population undergoes sex, which converts another 2% of the 97.7% remaining L into non-L (and 2% of the 2.3% non-L have undergone a 2nd meiosis). Etc. A more realistic model would remove synchrony, add stochasticity, and diatom bloom/bust population growth, but I think on average the result will be the same — exponential decline in the fraction of L cells in the population, something like exp(-ct/s) in the t/s "generations" imagined. [I still need to work out the equations more carefully, but note that exp(-200/10) = 2e-9.]

We've seen L in 5/7 tries, suggesting that L is a majority or prominent minority at 5/7 locations. To reconcile this with the above picture of exponentially decaying L population, it could be that:

1) L's age (in comparison to the average interval between sex) is so small that the exponential decline is still modest; I.e. t is << 200 or s >> 10
2) my scenario *assumed no fitness differences between L and its offspring.* But perhaps L is such a precariously balanced genotype that it out-competes all of its sexual offspring and/or they all have sex more often (and f < 1)
3) A variant of (2): (Ginger's devil's advocate proposal): L is really a tiny minority everywhere, but extremely adept at being brought into culture, and much more so than any of its sexual offspring
4) L is an obligate asexual (the extreme case of (1) where s=infinity).

To me, #4 is by far the most parsimonious explanation. L's global fitness is hard to explain, but nature trys trillions of genotypes, so presumably it can happen. Finding such a genotype that is in addition so much more fit than any of its sexual offspring (#2, #3) seems vastly harder. On the other hand, making a fit genotype asexual is essentially trivial - just break one of the many key genes involved in meiosis or gamete recognition.

As usual, my explanation is too long winded for the paper, and my list of counter-proposals is probably incomplete, but I think we're on pretty firm ground with the obligate asex hypothesis.

9/8/17, 2nd message:

PS: in case it wasn't obvious, the specific numbers I used in the example for concreteness (t=200, s=10, etc) were guesses at what I thought was a somewhat conservative extreme of young genotype, rare sex, but the specific values don't matter too much, I think we'll see the same qualitative behavior for any reasonable values.

9/8/17, 3rd message:

FYI, my first attempt at doing the algebra suggests the correct formula is exp(-ft/s), where f, t, and s are as above. E.g., using the example numbers from that message exp(-4)=0.018. But I certainly need to double-check it.

9/18, Feedback from Nao & my response:

> 1) Is it fair to say that the fitness of L-population is equal to that of non-L? Isn't a part of our argument that L-population is "general-purpose" population and that they better fit the general environment than the non-L population? Reviewers might question the fitness(L)=fitness(non-L) assumption.

L has a particular set of genes that make it fit (we think). The offspring of an L cross L mating will by and large have the same set of genes, so to first approximation at least, should be about equally fit. Ok, this is not totally accurate, because one allele of each gene is lost in newly homozygous regions, but it's just as likely to be a "bad" allele as a "good" allele that's lost. Note that my model isn't that L is equal to non-L; it's that L is equal to sexual offspring of L; it seems really unlikely to me that L should be more fit that all or most of its offspring.

[[9/29: above was poorly worded (I blame jet-lag), but hopefully clear. Reorder it to say "in newly hom regions, one allele is lost". I.e., if L has, say A/a genotype in some region, L x L cross may have A/A, a/a or A/a genotype. I.e., the later case retains the L-genotype, whereas the former two replace the mixed genotype with the pure "A" or pure "a" case. The "A/a" case is presumably neutral, while the other two cases are at worst a (random) mixture of beneficial and deleterious changes. We assume the overall net change to be neutral.]]

> 2) It seems little odd for me that entire L population suddenly becomes sexual. To me, more realistic starting point is to have a fraction of L population (e.g., 10%) reproducing at the rate of 10% per year? I might need to talk to Larry in person about this because I might be completely off.

If we're assumming L is genetically homogeneous, then all are sexual or all not. I think the reviewer's suggestion was that L has persisted because it has a much longer interval between sex, which is what my parameter "s" is about. If I remember, I think Julie told me s is less than 1 year for related species, not s=10 years; that was meant to be an extremely conservative choice.

9/19 exchange with Julie:

> Thinking about Nao's pt 2 below and Larry's response: If L goes sexual in an all or nothing manner because of genetic homogeneity - then L disappears on the first round of sex. Doesn't it? It's terminal commitment to gametogenesis. What AM I missing?

I mean genetically homogeneous but responding individually to whatever environmental clues trigger sex, and at whatever overall rate. Nao proposed 90% strictly asexual, and 10% sexual with mean interval s=10. An alternative is all sexual with mean interval s=100. These are different models, both with (initially) about 1%

of the pop undergoing meiosis in year 1. They have very different long term outcomes: in Nao's model, the sexual 10% will somewhat quickly (exp(-ft/10)) convert to non-L, leaving the 90% frozen in L-state, whereas in mine all will convert, but more slowly (exp(-ft/100). We can propose either, but homogeneous seems more parsimonious and I don't see a compelling reason to suspect that the few genetic differences among L-isolates include sex-related ones.