Progress, prospects, and practical suggestions for inferring demography and selection

with an application to human cytomegalovirus

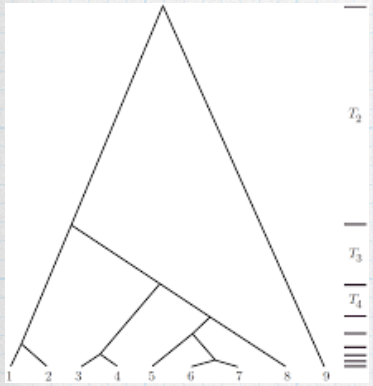
Jeffrey D. Jensen

June 13, 2016

jjensenlab.org

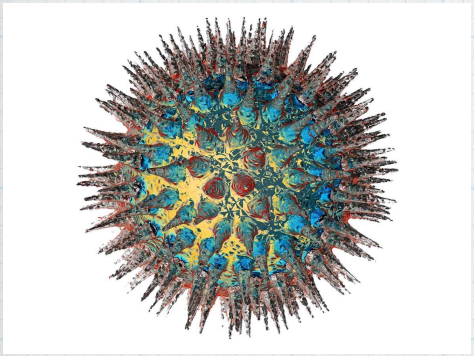
Outline

I. General introduction (~45 mins)



- the current state of population genetic inference
- pressing challenges
- future prospects

II. Research application (~15 mins)



- the population genetics of human cytomegalovirus (HCMV)
  - a direct application of the above suggested approaches
  - an illustration of the interplay between demography and selection
  - key clinical implications

Inferring the demographic and selective history of natural populations

What are the processes that we are certain are operating in natural populations?

- mutation (& recombination/reassortment, as applicable)
- genetic drift (and, relatedly, population size change and structure)
- purifying selection

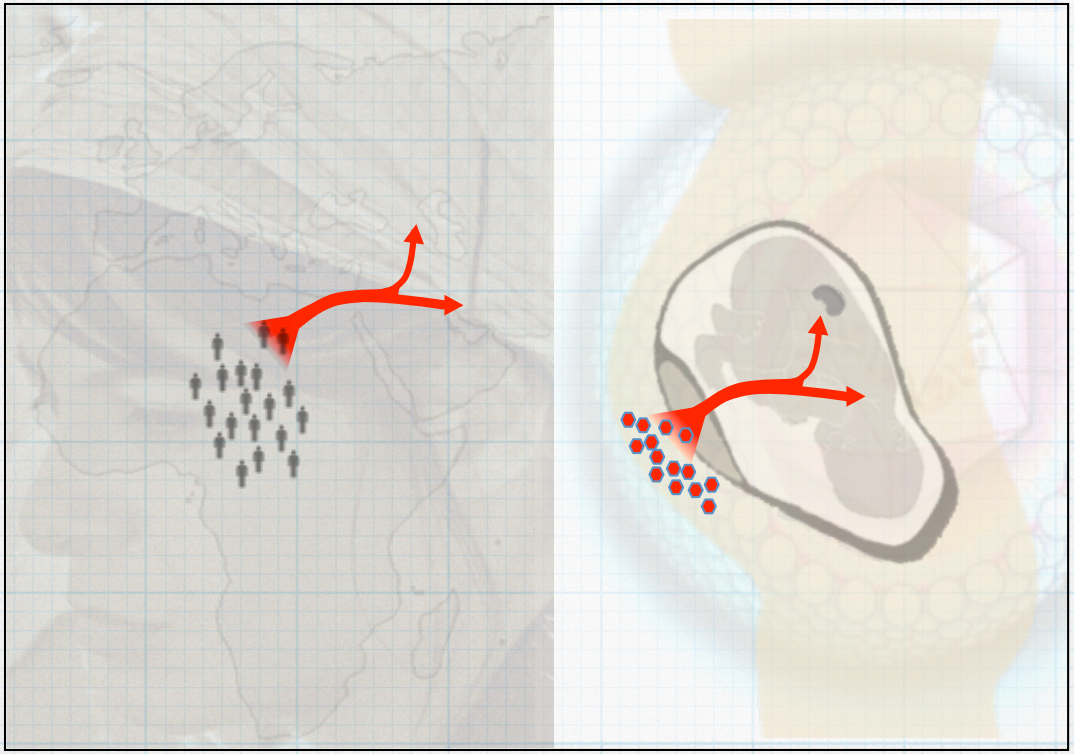
Inferring the demographic and selective history of natural populations

What are the processes that we are certain are operating in natural populations?

- mutation (& recombination/reassortment, as applicable)
- genetic drift (and, relatedly, population size change and structure)
- purifying selection

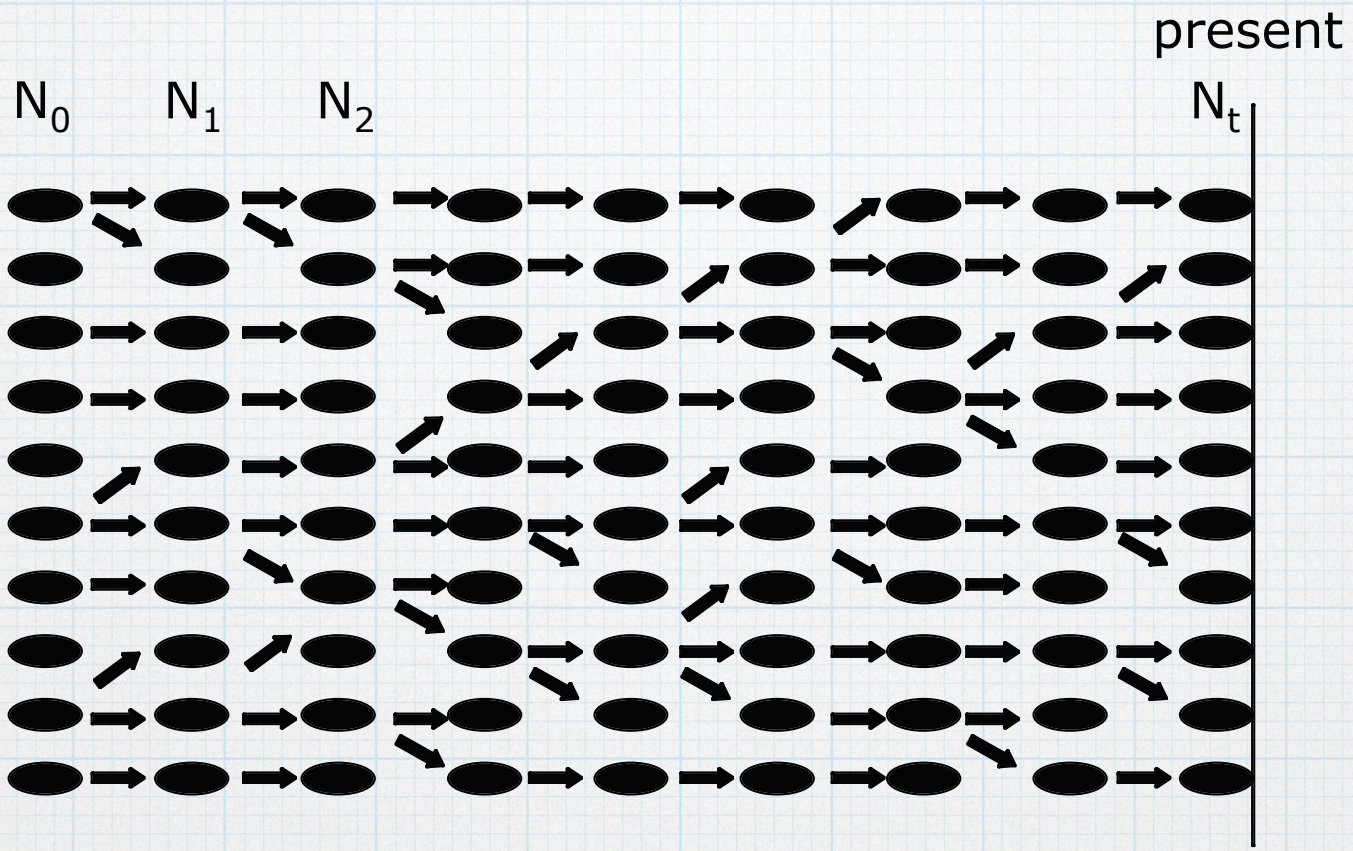
Thus, if we want to begin discussing inference, particularly with regards to adaptive evolution, we first need to understand these underlying effects - within the context of which positive selection may also be shaping genomic patterns

Inferring the demographic and selective history of natural populations



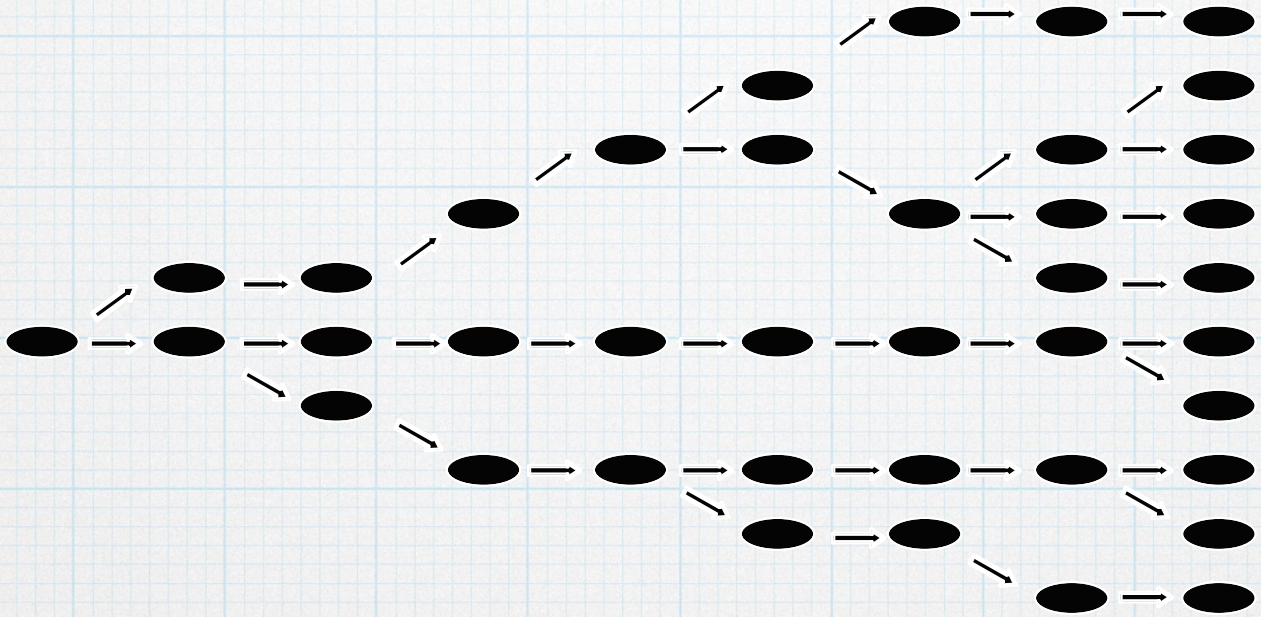
- For example, common demographic features include population size change (with population bottlenecks associated with colonization often being of particular interest), population structure, migration, and admixture

Let us begin by thinking about parents and offspring



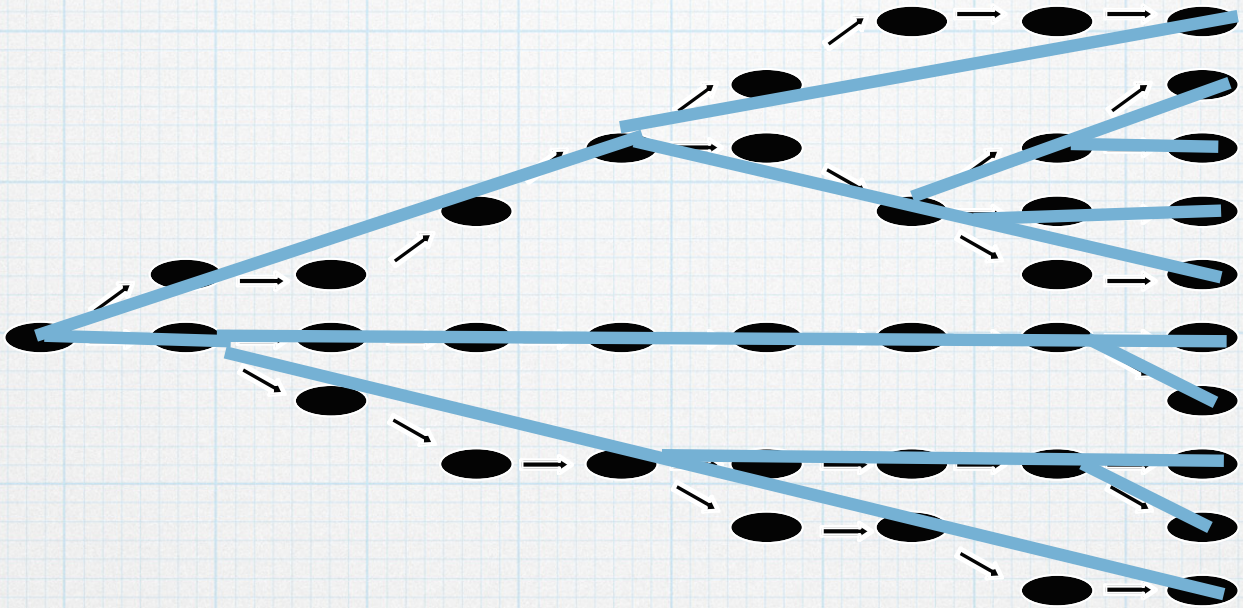
Considering a population of constant size over time, we see a process of sampling from one generation to the next

Coalescent events and a most recent common ancestor



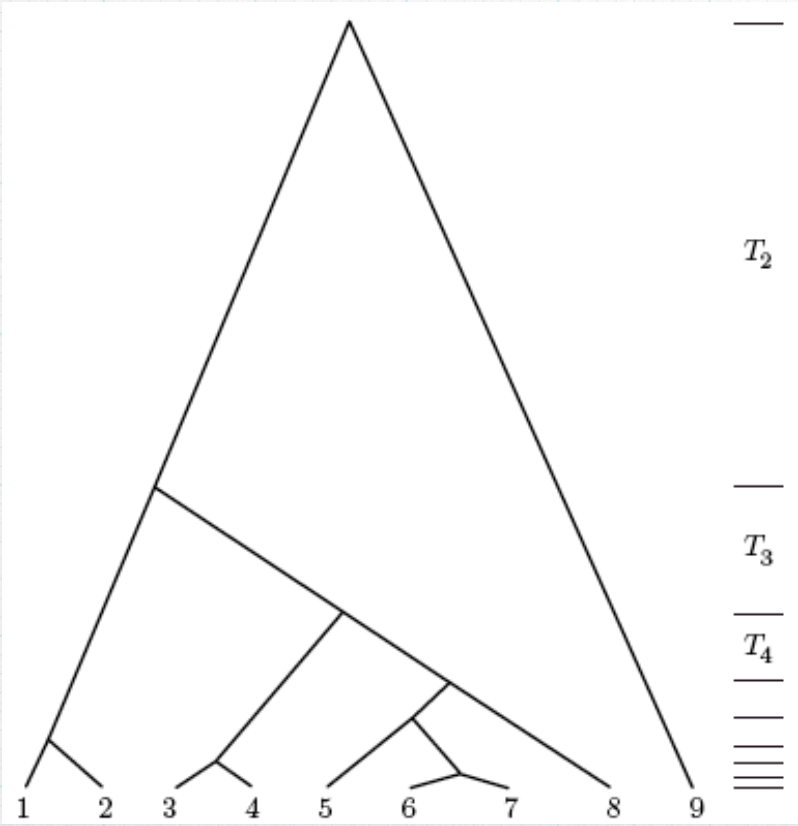
If we ignore individuals who did not contribute to current diversity, we easily see coalescent events and a most recent common ancestor

A coalescent tree



And let's just draw lines to connect these relationships - and there we have a coalescent tree

The standard neutral model



$$E[T_i] = \frac{2}{i(i-1)}$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

$$E[\xi_i] = \frac{\theta}{i}$$

Under a neutral model in which the population is at equilibrium, we understand a lot about the shape of this tree, and thus about expected patterns of variation

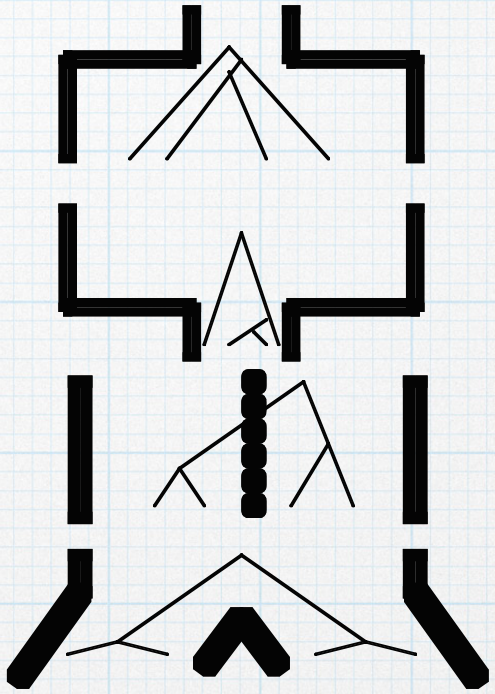
What happens when we relax the equilibrium assumption?

GROWTH

DECLINE

SUBDIVISION WITH MIGRATION

ISOLATION



adapted from J. Wakeley

With changing population size and structure, the shapes of these trees change owing to changes in the rate of coalescence. Thus, the expected patterns of variation in the population change as well.

So, how do we infer the underlying demographic history from these observed patterns of variation?

There are many existing approaches for estimating aspects of the demographic history of a population, assuming the sampling of neutral SNPs (i.e., not impacted by direct, or linked, selection)

So, how do we infer the underlying demographic history from these observed patterns of variation?

There are many existing approaches for estimating aspects of the demographic history of a population, assuming the sampling of neutral SNPs (i.e., not impacted by direct, or linked, selection)

- these take multiple statistical forms, but in principle all seek to fit a demographic model to observed patterns of variation

- and they thus all inherently are limited by model choice - that is, one may identify the best fitting demographic model of the models that the user has chosen to evaluate

Commonly used demographic estimators

In terms of commonly used approaches / software, here are a few of the most popular:

standard Approximate Bayesian Computation (ABC)
e.g., Thornton & Andolfatto 2006

- advantageous as user can define model, identify the best summary stats for said model, and readily test performance
- but is not a 'download and press button' approach, thus requires investment from user

Commonly used demographic estimators

In terms of commonly used approaches / software, here are a few of the most popular:

standard Approximate Bayesian Computation (ABC)
e.g., Thornton & Andolfatto 2006
- advantageous as user can define model, identify the best summary stats for said model, and readily test performance
- but is not a 'download and press button' approach, thus requires investment from user

dadi
Gutenkunst et al. 2009
- handles complex demography for up to three populations
- relatively slow and some convergence problems, but appears to perform very well

Commonly used demographic estimators

In terms of commonly used approaches / software, here are a few of the most popular:

standard Approximate Bayesian Computation (ABC)
e.g., Thornton & Andolfatto 2006
- advantageous as user can define model, identify the best summary stats for said model, and readily test performance
- but is not a 'download and press button' approach, thus requires investment from user

dadi
Gutenkunst et al. 2009
- handles complex demography for up to three populations
- relatively slow and some convergence problems, but appears to perform very well

fastsimcoal2
Excoffier et al. 2013
- can handle complex demographic models, and multiple populations
- simulations however suggest often poor fits to observed data

Commonly used demographic estimators

In terms of commonly used approaches / software, here are a few of the most popular:

standard Approximate Bayesian Computation (ABC)
e.g., Thornton & Andolfatto 2006
- advantageous as user can define model, identify the best summary stats for said model, and readily test performance
- but is not a 'download and press button' approach, thus requires investment from user

dadi
Gutenkunst et al. 2009
- handles complex demography for up to three populations
- relatively slow and some convergence problems, but appears to perform very well

fastsimcoal2
Excoffier et al. 2013
- can handle complex demographic models, and multiple populations
- simulations however suggest often poor fits to observed data

PSMC
Li & Durbin 2011
- estimates changes in *Ne* over time
- robustness to any diversity reducing effects is highly questionable

So, which is best to use?

Unfortunately, performance evaluation and comparison is something that not all reviewers are enforcing upon authors currently. Thus, it is difficult to know actually.

So, which is best to use?

Unfortunately, performance evaluation and comparison is something that not all reviewers are enforcing upon authors currently. Thus, it is difficult to know actually.

But I can tell you two things:

1) As you can easily simulate demographic models (in ms, msms, SFScode, SLiM, etc.), you can readily do this work yourself to evaluate which performs best for your given models and parameters of relevance for a given population

2) And just informally, our own simulations generally suggest that *dadi* is the best performing off-the-shelf method, but 'old-fashioned' ABC is actually the best and easiest way to tailor inference to your particular question.

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

3) utilization of multiple different approaches (we like dadi, fsc2, abc)

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

3) utilization of multiple different approaches (we like dadi, fsc2, abc)

4) simulation study to determine the power of these approaches for the models of interest

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

3) utilization of multiple different approaches (we like dadi, fsc2, abc)

4) simulation study to determine the power of these approaches for the models of interest

5) identification of the best fitting models and parameters from these approaches

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

3) utilization of multiple different approaches (we like dadi, fsc2, abc)

4) simulation study to determine the power of these approaches for the models of interest

5) identification of the best fitting models and parameters from these approaches

6) comparison of the fit of the best estimates from each to your data

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

3) utilization of multiple different approaches (we like dadi, fsc2, abc)

4) simulation study to determine the power of these approaches for the models of interest

5) identification of the best fitting models and parameters from these approaches

6) comparison of the fit of the best estimates from each to your data

7) simulation study to determine the ability of these estimators to accurately infer these best fitting parameters

What is the current best practice in demographic inference?

1) firstly, high quality SNPs, and mutation and recombination rate estimates
   - this could be a whole talk in itself, I suggest the recent review of Pfeifer for understanding the best practices for going from raw reads to usable population data

2) definition of models of relevance for your population, with a wide exploration

3) utilization of multiple different approaches (we like dadi, fsc2, abc)

4) simulation study to determine the power of these approaches for the models of interest

5) identification of the best fitting models and parameters from these approaches

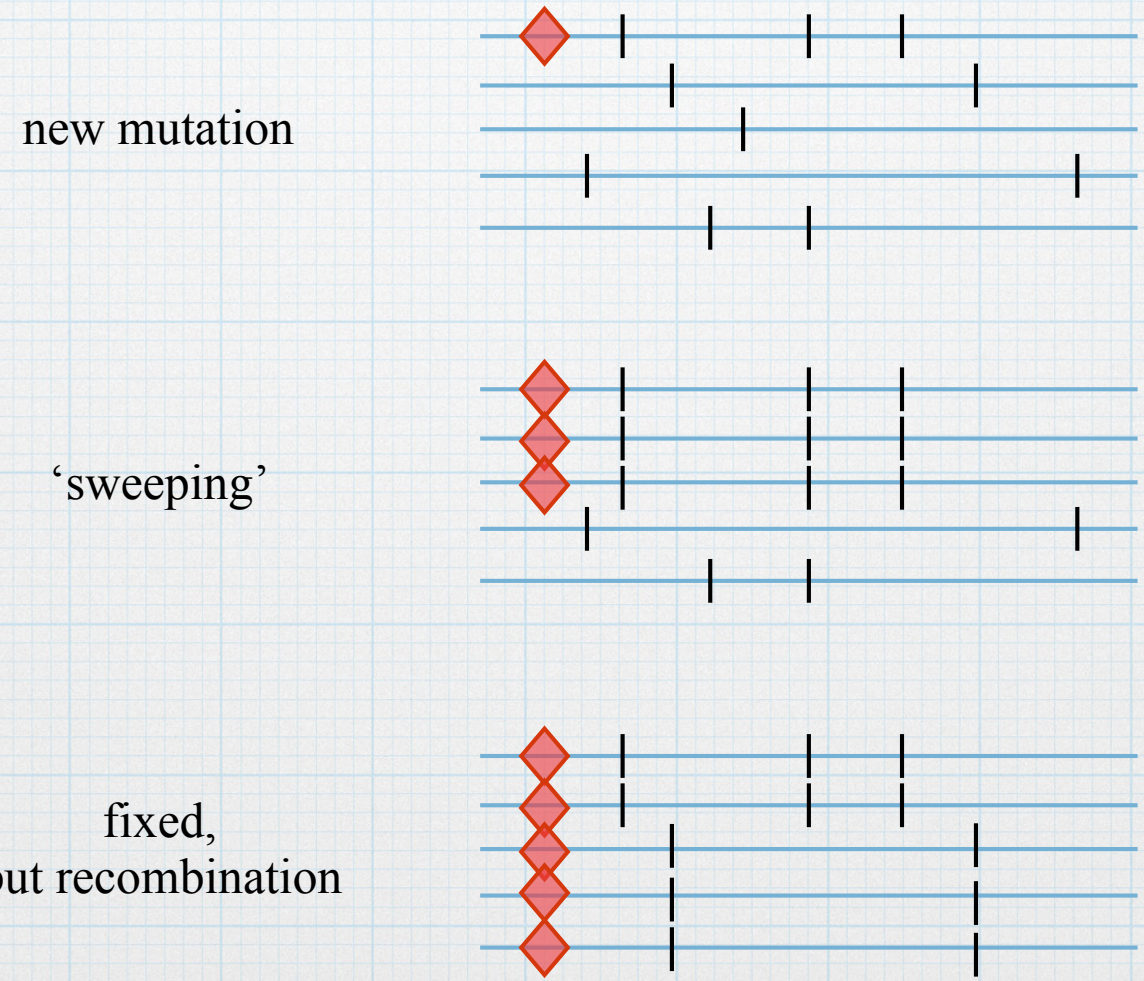
6) comparison of the fit of the best estimates from each to your data

7) simulation study to determine the ability of these estimators to accurately infer these best fitting parameters

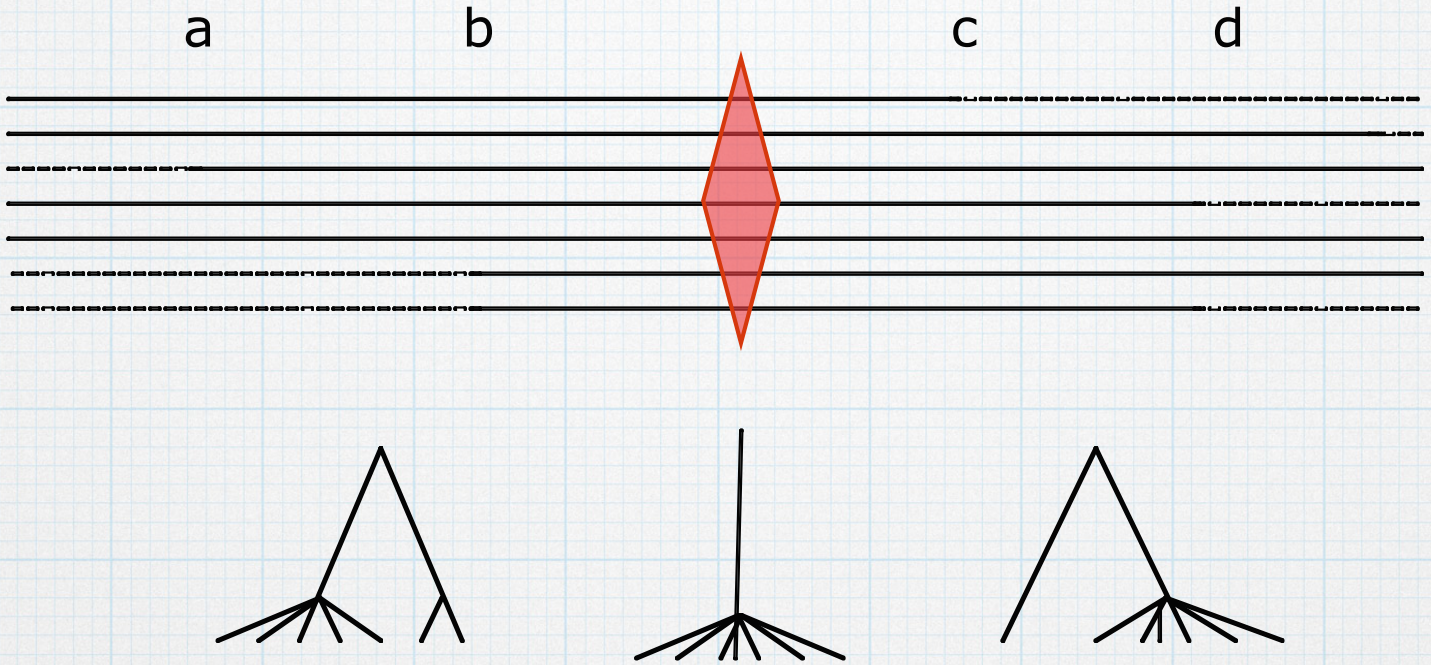
8) simulation study to determine the direction of bias of likely model violations (for example, for the effects of background selection)

So, as positive selection is occurring within the context of these non-equilibrium histories, how do we now map genetic hitchhiking events with these neutral demographic estimates in hand?

Firstly, what do we mean by genetic hitchhiking?



And what does this process do to the shape of the underlying coalescent trees?

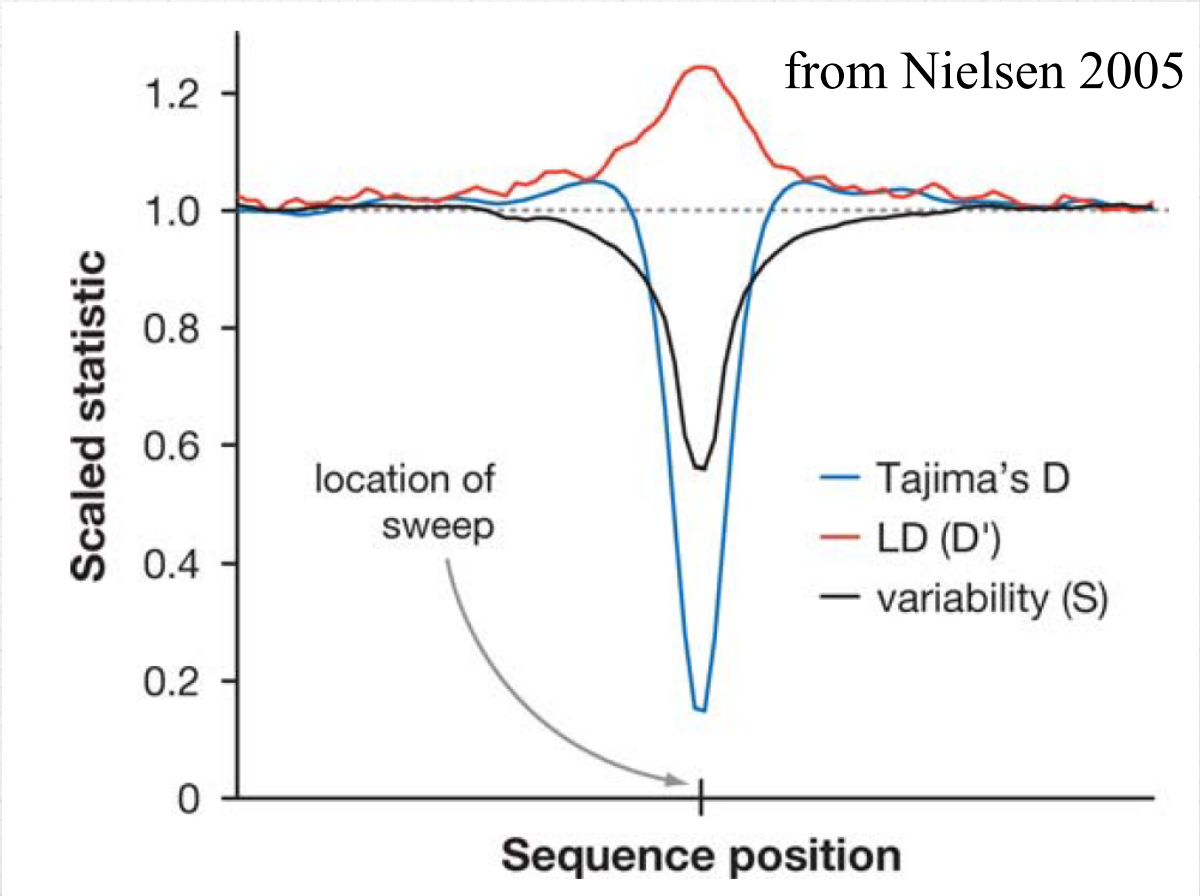


Thus, fundamentally, within a population genetic hitchhiking may alter the site frequency spectrum, patterns of linkage disequilibrium, and haplotype structure - relative to the standard neutal model

Current state of hitchhiking mapping

There are many existing approaches to map regions of the genome recently shaped by genetic hitchhiking, assuming the sampling of neutral SNPs (i.e., not impacted by direct selection, but indeed impacted by linked selection)

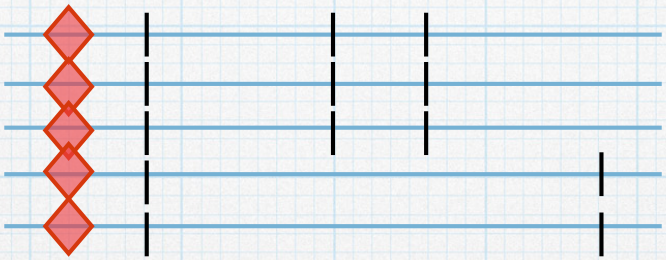
- these generally use one or more of the following:
  - the level of variation
  - the shape of the site frequency spectrum
  - linkage disequilibrium and haplotype structure
  - differentiation between populations



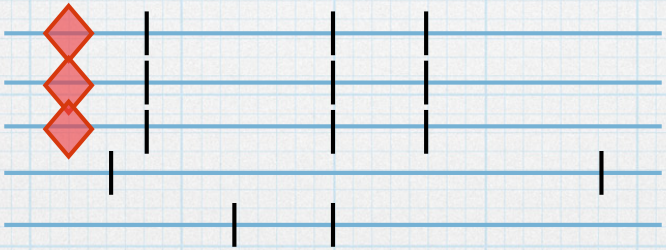
Hitchhiking models of interest

In general, there are a handful of models of interest to people, and different approaches are used for each as expected patterns of variation differ

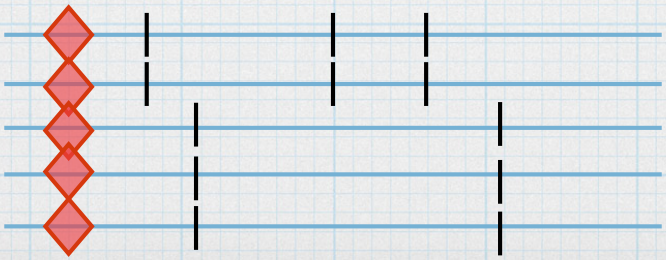
- a complete selective sweep from a newly arising mutation



- an incomplete selective sweep



- a complete selective sweep from a previously segregating variant



Existing approaches for detecting hitchhiking patterns

In terms of commonly used approaches / software, here are a few of the most popular:

Composite Likelihood Ratio (CLR) and CLR-based (including Sweepfinder and such)
- fundamentally from Kim&Stephan 2002, but many many subsequent extensions
- calculates the CL (thus SFS based) of the data under a hard sweep model vs a neutral model

Existing approaches for detecting hitchhiking patterns

In terms of commonly used approaches / software, here are a few of the most popular:

Composite Likelihood Ratio (CLR) and CLR-based (including Sweepfinder and such)
- fundamentally from Kim&Stephan 2002, but many many subsequent extensions
- calculates the CL (thus SFS based) of the data under a hard sweep model vs a neutral model

omega
from Kim&Nielsen 2004 and Jensen et al. 2007
- similar to above statistically, though evalutes linkage disequilibrium based expectation

Existing approaches for detecting hitchhiking patterns

In terms of commonly used approaches / software, here are a few of the most popular:

Composite Likelihood Ratio (CLR) and CLR-based (including Sweepfinder and such)
- fundamentally from Kim&Stephan 2002, but many many subsequent extensions
- calculates the CL (thus SFS based) of the data under a hard sweep model vs a neutral model

omega
from Kim&Nielsen 2004 and Jensen et al. 2007
- similar to above statistically, though evalutes linkage disequilibrium based expectation

EHH/iHS
from Sabeti et al. 2002 and Voight et al. 2006
- looks for long stretches of haplotype structure to identify incomplete sweeps

Existing approaches for detecting hitchhiking patterns

In terms of commonly used approaches / software, here are a few of the most popular:

Composite Likelihood Ratio (CLR) and CLR-based (including Sweepfinder and such)
- fundamentally from Kim&Stephan 2002, but many many subsequent extensions
- calculates the CL (thus SFS based) of the data under a hard sweep model vs a neutral model

omega
from Kim&Nielsen 2004 and Jensen et al. 2007
- similar to above statistically, though evalutes linkage disequilibrium based expectation

EHH/iHS
from Sabeti et al. 2002 and Voight et al. 2006
- looks for long stretches of haplotype structure to identify incomplete sweeps

xp-EHH, xp-CLR, Bayescan
from Sabeti et al. 2007, Chen et al. 2010, and Foll&Gaggiotti 2008
- taking the above EHH or CLR-based statistics between rather than within populations, as well as looking for Fst outliers

So, which is best to use?

As with demographic inference, testing between methods is unfortunately not yet a fully enforced standard in the field. However, more is known here, particularly as the concern about mis-inference under demographic models has been appreciated for some time.

So, which is best to use?

As with demographic inference, testing between methods is unfortunately not yet a fully enforced standard in the field. However, more is known here, particularly as the concern about mis-inference under demographic models has been appreciated for some time.

The first thing to consider is how these approaches perform under ideal conditions (i.e., equilibrium population histories)

Table 5 | True positive rate for SHH selection models.

2Ns	10	100	1000
SweepFinder	0.05	0.14	0.33
SweeD	0.05	0.13	0.32
SweeD with monomorphic	0.12	0.15	0.34
OmegaPlus	0.07	0.37	0.46

from Crisci et al. 2013, *Frontiers*

So, which is best to use?

The next question you want to ask, is the true positive and false positive rate under common non-equilibrium models.

Table 3 | False positive rate for neutral bottleneck models (sfscode).

Reduction (%)	Age of bottleneck recovery event (2*N*generations)														
	0.02					0.05					0.1				
	25	50	75	90	99	25	50	75	90	99	25	50	75	90	99
SweepFinder	0.01	0.02	0.08	0.08	0.01	0.02	0.03	0.04	0.05	0.01	0.01	0.01	0.04	0.04	0.00
SweeD	0.07	0.09	0.13	0.07	0.01	0.06	0.09	0.09	0.07	0.00	0.05	0.09	0.11	0.08	0.00
SweeD with monomorphic	0.13	0.19	0.27	0.18	0.01	0.15	0.18	0.17	0.16	0.00	0.12	0.16	0.13	0.11	0.00
OmegaPlus	0.07	0.13	0.26	0.31	0.68	0.09	0.13	0.26	0.34	0.91	0.08	0.12	0.22	0.43	0.79

from Crisci et al. 2013, *Frontiers*

So, which is best to use?

The next question you want to ask, is the true positive and false positive rate under common non-equilibrium models.

Table 3 | False positive rate for neutral bottleneck models (sfscode).

Reduction (%)	Age of bottleneck recovery event (2*N*generations)														
	0.02					0.05					0.1				
	25	50	75	90	99	25	50	75	90	99	25	50	75	90	99
SweepFinder	0.01	0.02	0.08	0.08	0.01	0.02	0.03	0.04	0.05	0.01	0.01	0.01	0.04	0.04	0.00
SweeD	0.07	0.09	0.13	0.07	0.01	0.06	0.09	0.09	0.07	0.00	0.05	0.09	0.11	0.08	0.00
SweeD with monomorphic	0.13	0.19	0.27	0.18	0.01	0.15	0.18	0.17	0.16	0.00	0.12	0.16	0.13	0.11	0.00
OmegaPlus	0.07	0.13	0.26	0.31	0.68	0.09	0.13	0.26	0.34	0.91	0.08	0.12	0.22	0.43	0.79

Table 6 | True positive rate for joint SHH-bottleneck models.

Reduction (%)	Age of bottleneck recovery event (2*N*generations)								
	0.02			0.05			0.1		
	50	25	90	50	25	90	50	25	90
***2Ns* = 100**									
SweepFinder	0.11	0.14	0.08	0.10	0.14	0.06	0.08	0.10	0.07
SweeD	0.09	0.12	0.07	0.09	0.13	0.05	0.08	0.09	0.07
SweeD with monomorphic	0.18	0.26	0.17	0.20	0.21	0.14	0.15	0.14	0.10
OmegaPlus	0.09	0.29	0.32	0.15	0.26	0.41	0.12	0.24	0.51
***2Ns* = 1000**									
SweepFinder	0.13	0.12	0.07	0.10	0.16	0.07	0.07	0.10	0.06
SweeD	0.12	0.11	0.07	0.09	0.13	0.06	0.07	0.09	0.05
SweeD with monomorphic	0.19	0.25	0.18	0.19	0.23	0.16	0.15	0.14	0.09
OmegaPlus	0.13	0.29	0.31	0.14	0.25	0.40	0.11	0.23	0.47

Note some statistics, like Sweepfinder, simply have no power to reject under these models.

Others (like Omega) do have power, but it is associated with a high FPR under some models.

from Crisci et al. 2013, *Frontiers*

What is the current best practice in hitchhiking mapping?

1) firstly, high quality SNPs, and mutation and recombination rate estimates (as described above)

What is the current best practice in hitchhiking mapping?

1) firstly, high quality SNPs, and mutation and recombination rate estimates (as described above)

2) demographic inference for the population of interest (as described above)

What is the current best practice in hitchhiking mapping?

1) firstly, high quality SNPs, and mutation and recombination rate estimates (as described above)

2) demographic inference for the population of interest (as described above)

3) simulation study to determine the power of existing statistics to identify and quantify hitchhiking effects under this demographic model, and the false positive rate

What is the current best practice in hitchhiking mapping?

1) firstly, high quality SNPs, and mutation and recombination rate estimates (as described above)

2) demographic inference for the population of interest (as described above)

3) simulation study to determine the power of existing statistics to identify and quantify hitchhiking effects under this demographic model, and the false positive rate

4) application of both SFS, LD/haplotype, and differentiation based approaches - as they have different power in different parameter spaces

What is the current best practice in hitchhiking mapping?

1) firstly, high quality SNPs, and mutation and recombination rate estimates (as described above)

2) demographic inference for the population of interest (as described above)

3) simulation study to determine the power of existing statistics to identify and quantify hitchhiking effects under this demographic model, and the false positive rate

4) application of both SFS, LD/haplotype, and differentiation based approaches - as they have different power in different parameter spaces

5) simulation study to determine the fit of the estimated parameters within the given demographic model to replicate observed patterns of variation

Right, so why is differentiating positive selection and demography so difficult?

Well, fundamentally, positive selection is like a population reduction followed by growth.

- that is, there are a disproportionate number of coalescent events at a particular point in time, followed by growth (i.e., a star-shaped tree)

- this was already well pointed out by Barton 1998

But, you may be thinking, positive selection has local effects and demography has genome wide effects

Its not that simple.

- One key issue is that a population bottleneck for example inflates the variance across the genome. Thus, there may appear to be localized effects, but simply because the tails of these distributions are now more extreme.

But, you may be thinking, positive selection has local effects and demography has genome wide effects

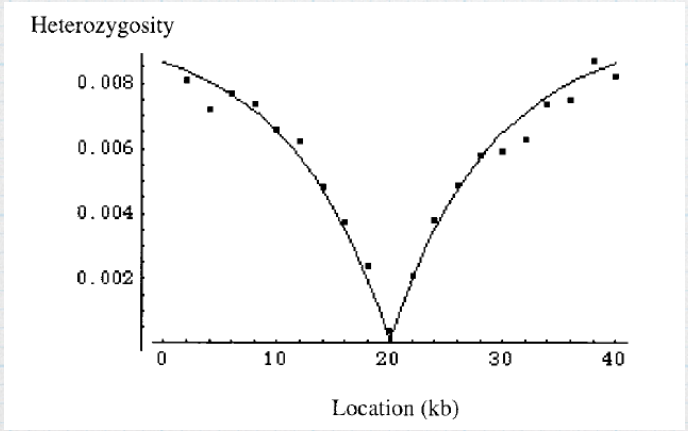
Consider practically how a genome scan works. A researcher may scan for a hitchhiking-like prediction (reduced variation, skewed frequency spectrum, elevated LD, etc.), and then focus in on that region (often with further sequencing) to apply tests of selection.

But, you may be thinking, positive selection has local effects and demography has genome wide effects

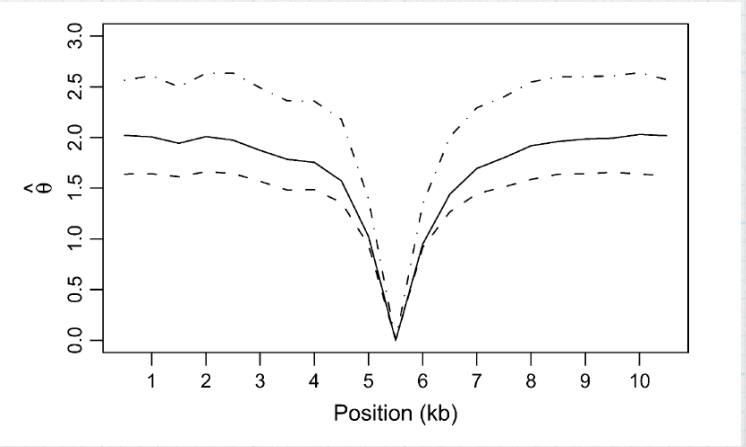
Consider practically how a genome scan works. A researcher may scan for a hitchhiking-like prediction (reduced variation, skewed frequency spectrum, elevated LD, etc.), and then focus in on that region (often with further sequencing) to apply tests of selection.

However, this creates a strong ascertainment bias. Specifically, the region has been ascertained for having hitchhiking-like patterns of variation, and when not subsequently accounted for under a proper demographic model, this leads to outrageously high false positive rates.

Expected hitchhiking pattern, from Kim&Stephan 2002



Expected pattern of ascertainment in a bottleneck, from Thornton&Jensen 2007



What we know about our ability to differentiate these processes?

1. We know that a large range of neutral demographic models replicate many patterns of variation expected under genetic hitchhiking models, potentially leading to very high false positive rates.

2. On the other side of the coin, genetic hitchhiking effects may be well fit by neutral demographic models.

What we know about our ability to differentiate these processes?

1. We know that a large range of neutral demographic models replicate many patterns of variation expected under genetic hitchhiking models, potentially leading to very high false positive rates.

2. On the other side of the coin, genetic hitchhiking effects may be well fit by neutral demographic models.

And herein lies the crux of the problem:

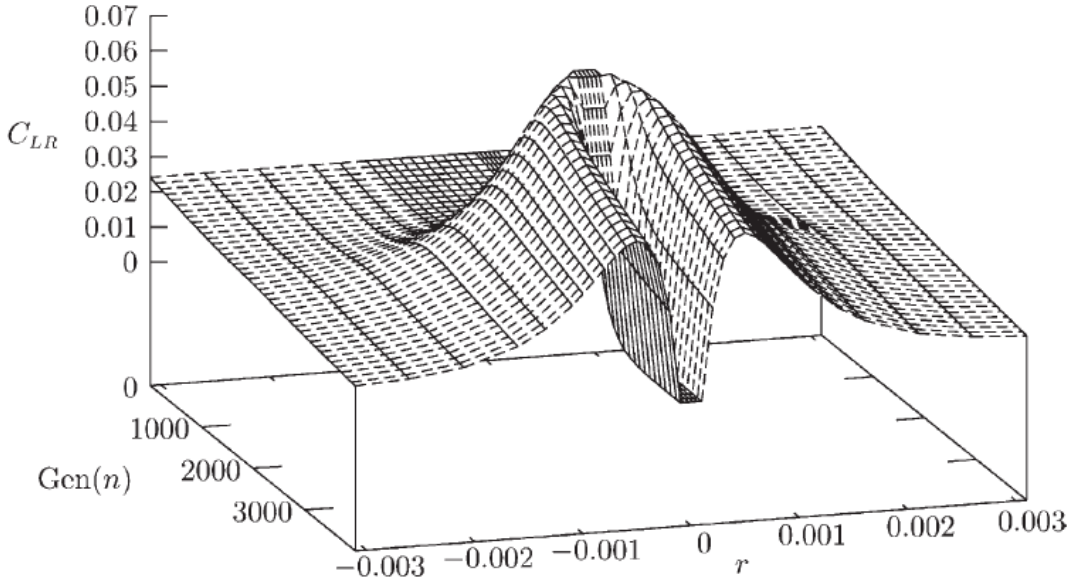
- by assuming neutrality in demographic estimation, one may be overfitting a demographic model to account for the localized effects of genetic hitchhiking (e.g., by estimating an overly severe bottleneck)

- and by neglecting demography, one may be 'identifying' regions under genetic hitchhiking that are simply the tails of neutral non-equilibrium effects

How have people proposed to enable this differentiation?

Basic ideas:

1) identify patterns of variation that are uniquely produced by genetic hitchhiking



from Stephan et al. 2006

- the best within population patterns seem to be the linkage disequilibrium effects associated with the omega statistics
  - -however, severe bottleneck models can indeed replicate it, and this pattern is very short-lived after a fixation event

How have people proposed to enable this differentiation?

Basic ideas:

2) discount the effects of positive selection, and focus on fitting data to demographic models

- to which I am sympathetic, as we know that non-equilibrium demography is indeed relevant to natural population
- but ignoring the effects of background selection is probably highly problematic here (more on this in a moment)

How have people proposed to enable this differentiation?

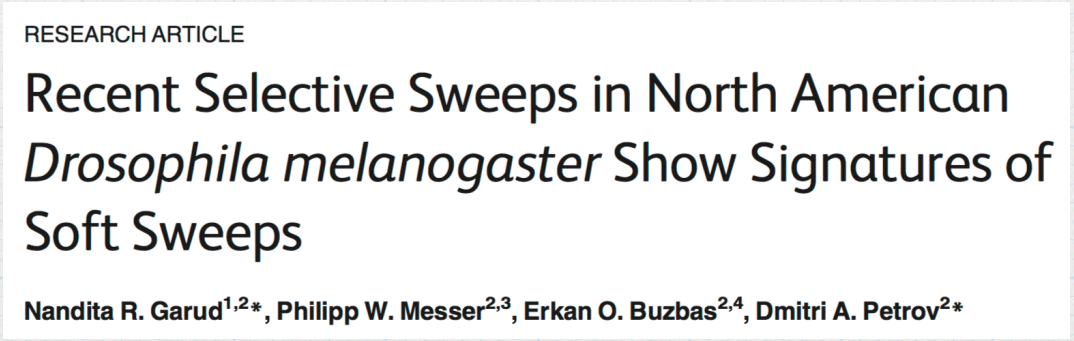
Basic ideas:

2) discount the effects of positive selection, and focus on fitting data to demographic models

- to which I am sympathetic, as we know that non-equilibrium demography is indeed relevant to natural population
- but ignoring the effects of background selection is probably highly problematic here (more on this in a moment)

3) discount the effects of demography, and focus on fitting data to postive selection models

- to which I am not at all sympathetic…
- to see the real danger of such poor population genetics, see most of Petrov's work for examples



RESEARCH ARTICLE

Recent Selective Sweeps in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps

Nandita R. Garud[1,2]*, Philipp W. Messer[2,3], Erkan O. Buzbas[2,4], Dmitri A. Petrov[2]*

How have people proposed to enable this differentiation?

Basic ideas:

4) do not attempt to fit a demographic model, but simply quantify the background levels of variation and find outliers (i.e., Sweepfinder, EHH, etc.)

- this approach is very dangerous, as you have no sense of the variance produced by the underlying demographic model
- and simulation results suggest that they do not work well as we saw above

Table 6 | True positive rate for joint SHH-bottleneck models.

Reduction (%)	Age of bottleneck recovery event ($2N$generations)								
	0.02			0.05			0.1		
	50	25	90	50	25	90	50	25	90
$2Ns = 100$									
SweepFinder	0.11	0.14	0.08	0.10	0.14	0.06	0.08	0.10	0.07

How have people proposed to enable this differentiation?

Basic ideas:

4) do not attempt to fit a demographic model, but simply quantify the background levels of variation and find outliers (i.e., Sweepfinder, EHH, etc.)

- this approach is very dangerous, as you have no sense of the variance produced by the underlying demographic model
- and simulation results suggest that they do not work well as we saw above

Table 6 | True positive rate for joint SHH-bottleneck models.

Reduction (%)	Age of bottleneck recovery event ($2N$generations)								
	0.02			0.05			0.1		
	50	25	90	50	25	90	50	25	90
$2Ns = 100$									
SweepFinder	0.11	0.14	0.08	0.10	0.14	0.06	0.08	0.10	0.07

5) attempt to jointly estimate parameters of demography and positive selection

- attempts have been made by Li&Stephan and others, but these approaches are generally step-wise and thus have the same problems outlined above, and have not proven very successful yet
- that being said, this is the right direction to focus future study

So, in practice, whats the current state of the art?

- 1) Estimate demographic model using intergenic SNPs following the 8 steps outlined above

So, in practice, whats the current state of the art?

- 1) Estimate demographic model using intergenic SNPs following the 8 steps outlined above

- 2) Simulate to characterize false positive rate and power of tests of selection under this demographic model

So, in practice, whats the current state of the art?

- 1) Estimate demographic model using intergenic SNPs following the 8 steps outlined above

- 2) Simulate to characterize false positive rate and power of tests of selection under this demographic model

- 3) Identify candidate SNPs using the best performing approaches for this parameter space identified in Step 2 (if Step 2 indeed suggests identification is possible)

So, in practice, whats the current state of the art?

- 1) Estimate demographic model using intergenic SNPs following the 8 steps outlined above

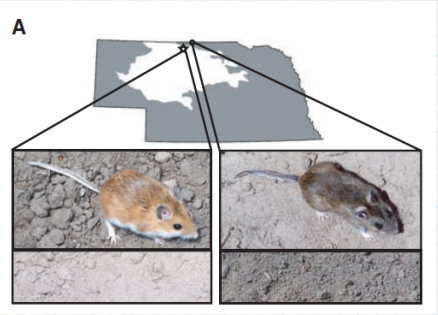
- 2) Simulate to characterize false positive rate and power of tests of selection under this demographic model

- 3) Identify candidate SNPs using the best performing approaches for this parameter space identified in Step 2 (if Step 2 indeed suggests identification is possible)

- 4) Bring SNPs to functional validation

A brief example of the application of these steps to some of our work

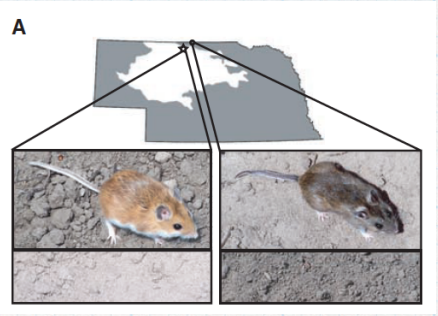
- 1) Estimate the demographic history, and simulate to characterize power to estimate the model, the fit of the model to the data, and power and false positive rate to detect selection within that model



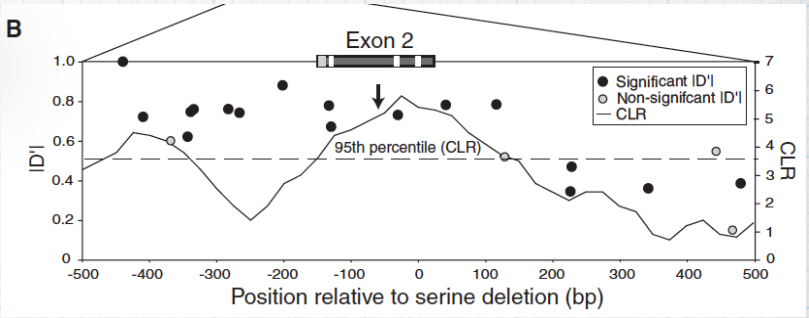
- Linnen et al. 2009, *Science*
- Linnen et al. 2013, *Science*

A brief example of the application of these steps to some of our work

- 1) Estimate the demographic history, and simulate to characterize power to estimate the model, the fit of the model to the data, and power and false positive rate to detect selection within that model



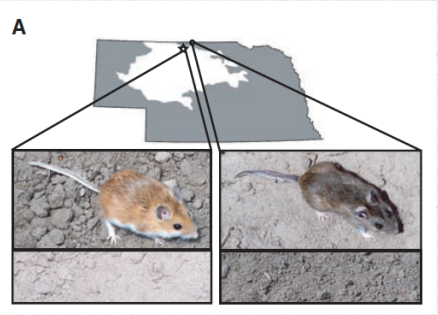
- 2) Identify putative sites under selection, and simulate to quantify ability to infer selection parameters under the demographic model



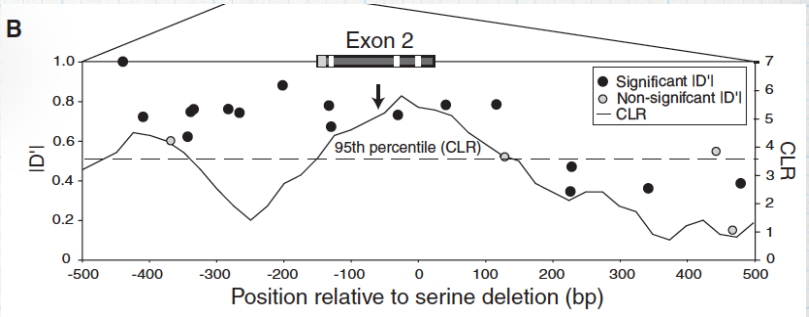
- Linnen et al. 2009, *Science*
- Linnen et al. 2013, *Science*

A brief example of the application of these steps to some of our work

- 1) Estimate the demographic history, and simulate to characterize power to estimate the model, the fit of the model to the data, and power and false positive rate to detect selection within that model



- 2) Identify putative sites under selection, and simulate to quantify ability to infer selection parameters under the demographic model



- 3) Bring identified SNPs to functional validation

S. Pfeifer

see Susanne's poster for more details about this work

- Linnen et al. 2009, *Science*
- Linnen et al. 2013, *Science*

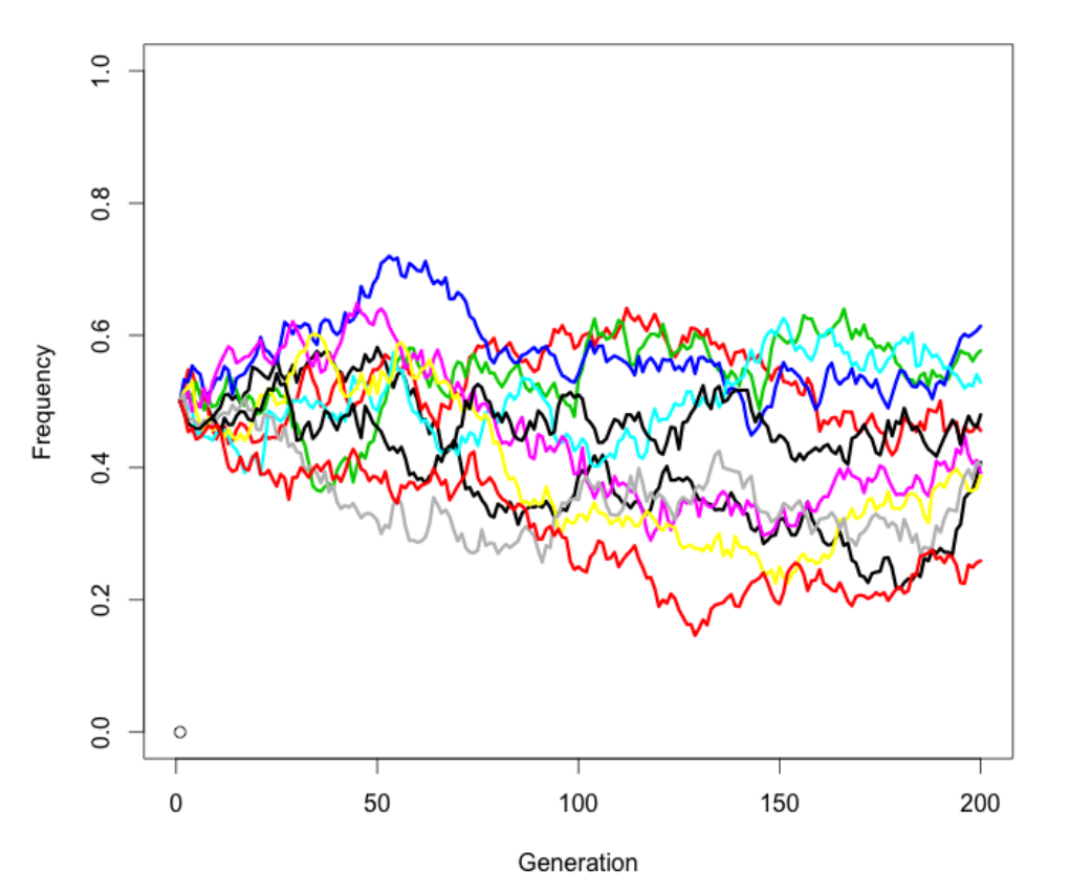
A game-changer, the recent addition of time-sampled polymorphism data

- While the above related my perception of the current best practices in analyzing 'standard' polymorphism data (that is, collected at a single time point), an important innovation is now common in multiple organisms - that is, the generation of polymorphism data sampled at multiple time points

    - such datasets are extremely common in experimental evolution and clinically relevant studies, increasingly common in ecological studies as well, and have a natural relation to the growing field of ancient DNA

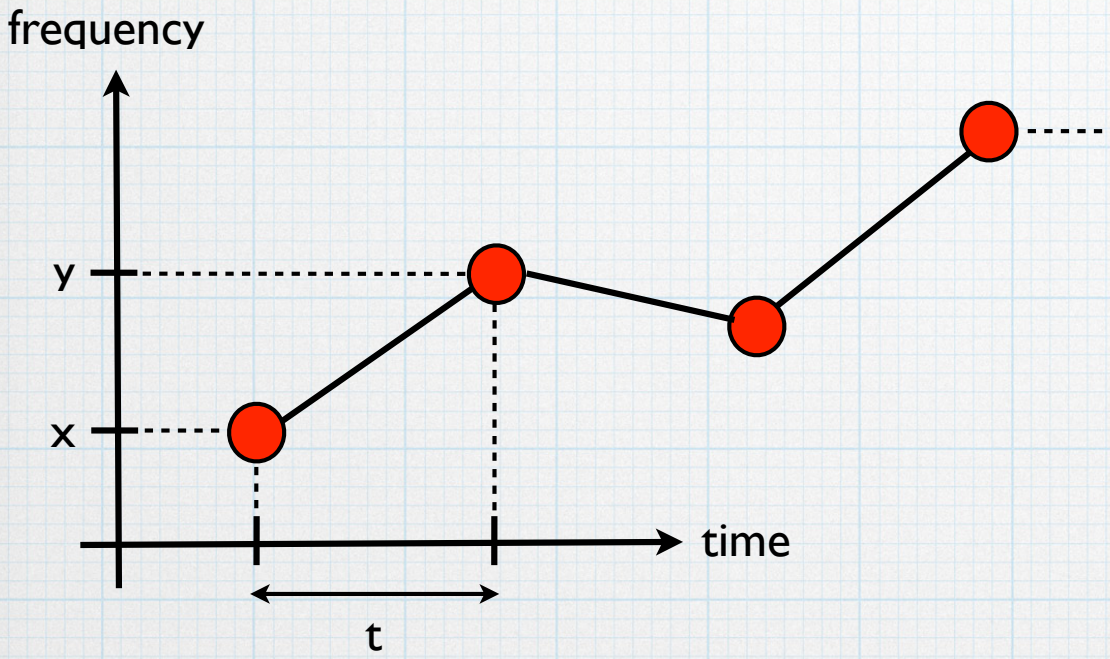
- this in principle allows for the estimation of selection coefficients, and effective population sizes, directly from the trajectories of each individual allele observed

So, fundamentally, how do such methods work?



- by tracking allele trajectories through time, one can calculate effective population size (*Ne*) from the per generation variance in frequencies
    - with that, one can ask which (if any) sites in the genome are changing too fast to be consistent with that *Ne*, and what selection coefficient is necessary to explain the magnitude of change

How this temporal dimension is helping to better differentiate selection and demography



- thus, rather than a single time-point realization of the site frequency spectrum of linked SNPs as we have for standard polymorphism data, we can here make inference through time by directly tracking the sites themselves that may be targeted by selection

Such methodologies are proliferating very rapidly in the literature

Here's what we know about a handful of such approaches: (again, the field is suffering from a lack of performance testing and comparison)

2008	Bollback…. Nielsen	ML	Severely biased. Never works. Ever.
2012	Malaspinas…Slatkin	ML	Very accurate for small *s*. Computationally intensive.
2013	Mathieson…McVean	ML	Fast, but *Ne* must be known.
2014	Foll…. Jensen	ABC	Fast, more accurate for large *s*. *Ne* estimate assumes mostly neutral loci.
2016	Shim….Jensen	ABC	Same as Foll et al., but detects and estimates changing *s* through time
2016	Ferrer….Wegmann	Bayesian	Faster approx. of WF diffusion, but still slow and thus only candidate sites

A brief example of the application of such work to the interests of our lab

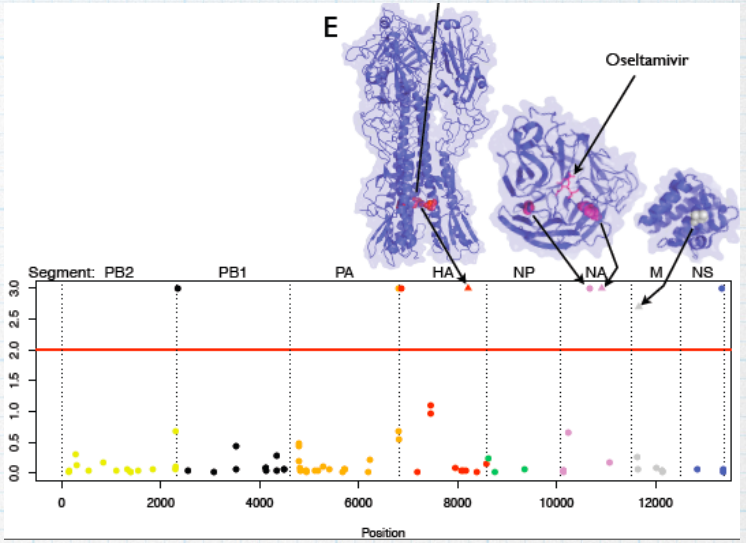
- 1) We're interested in using these approaches on experimentally passaged populations of influenza virus, in order to evaluate both new and existing drugs, individually and in combination

- Foll et al. 2014, *PLOS Genetics*
- Matuszewski *et al., Genetics,* in press
- Bank et al., in review

A brief example of the application of such work to the interests of our lab

- 1) We're interested in using these approaches on experimentally passaged populations of influenza virus, in order to evaluate both new and existing drugs, individually and in combination

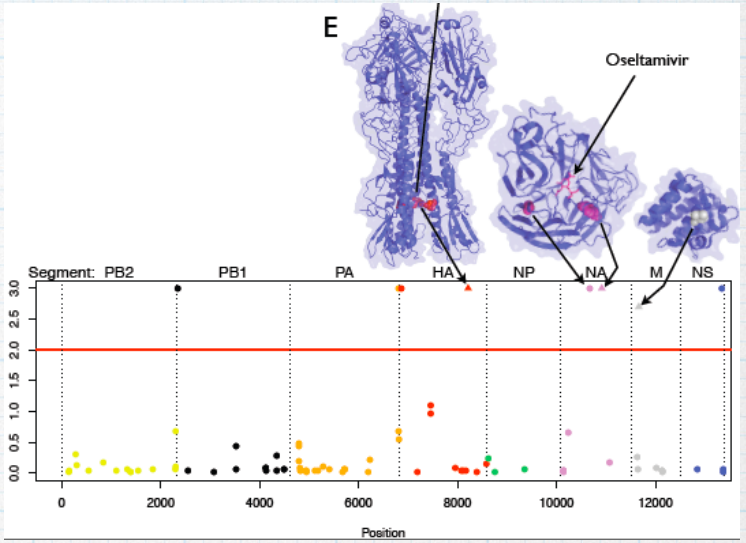
- 2) In the case of the most common drug treatment (oseltamivir), we've shown easily accessible mutational routes to resistance



- Foll et al. 2014, *PLOS Genetics*
- Matuszewski *et al., Genetics,* in press
- Bank et al., in review

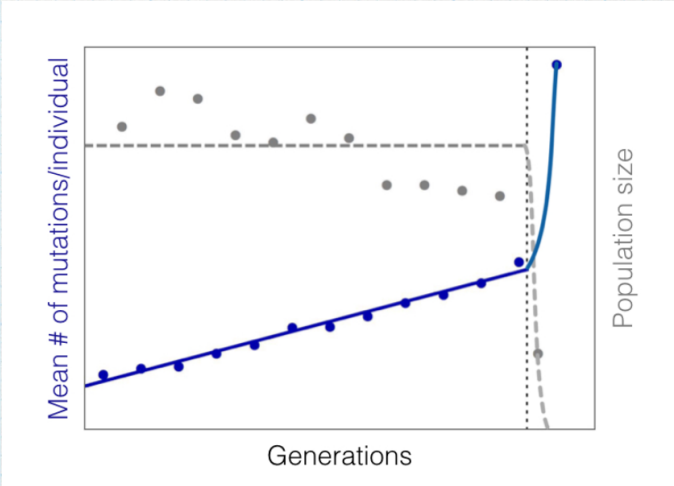
A brief example of the application of such work to the interests of our lab

- 1) We're interested in using these approaches on experimentally passaged populations of influenza virus, in order to evaluate both new and existing drugs, individually and in combination



- 2) In the case of the most common drug treatment (oseltamivir), we've shown easily accessible mutational routes to resistance

- 3) Conversely, in an experimental drug (favipiravir) which increases viral mutation rate, we've recently shown that it is possible to achieve mutational meltdown



S. Matuszewski

see Sebastian's poster for more details about this work

- Foll et al. 2014, *PLOS Genetics*
- Matuszewski *et al., Genetics,* in press
- Bank et al., in review

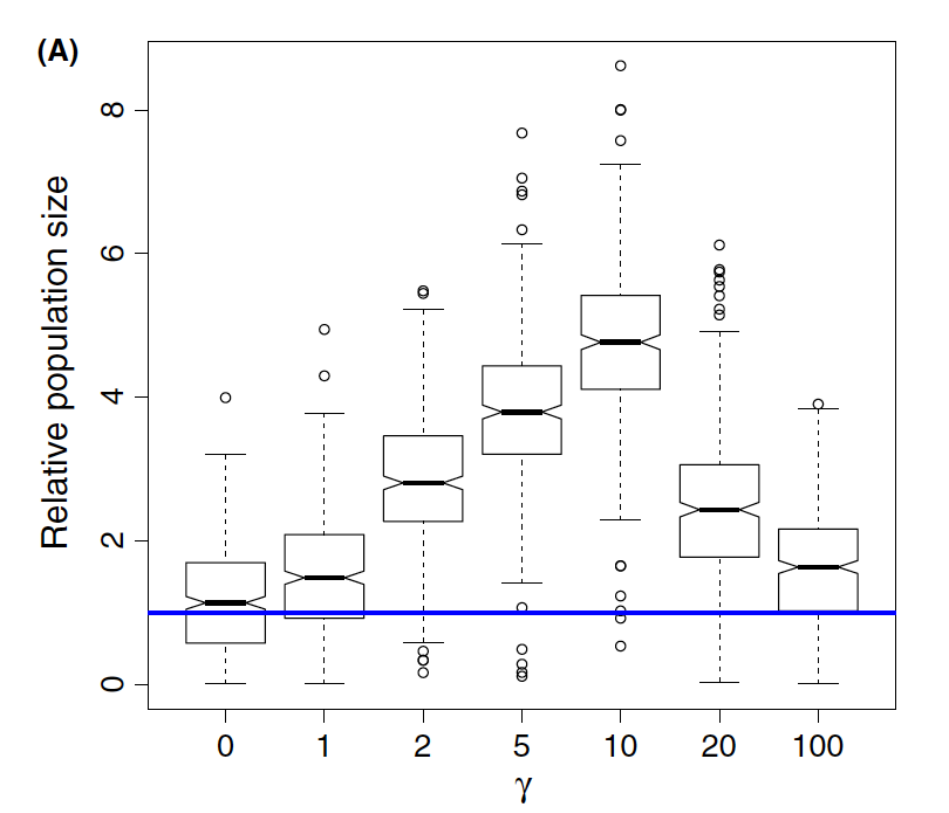
So, to wrap up this section: Outstanding Issue #1 in population genetic inference: incorporating the effects of BGS

- **Purifying selection is the dominant and pervasive mode of selection across the genome, and the resulting effects of background selection (BGS) shape the site frequency spectrum**

So, to wrap up this section: Outstanding Issue #1 in population genetic inference: incorporating the effects of BGS

- **Purifying selection is the dominant and pervasive mode of selection across the genome, and the resulting effects of background selection (BGS) shape the site frequency spectrum**

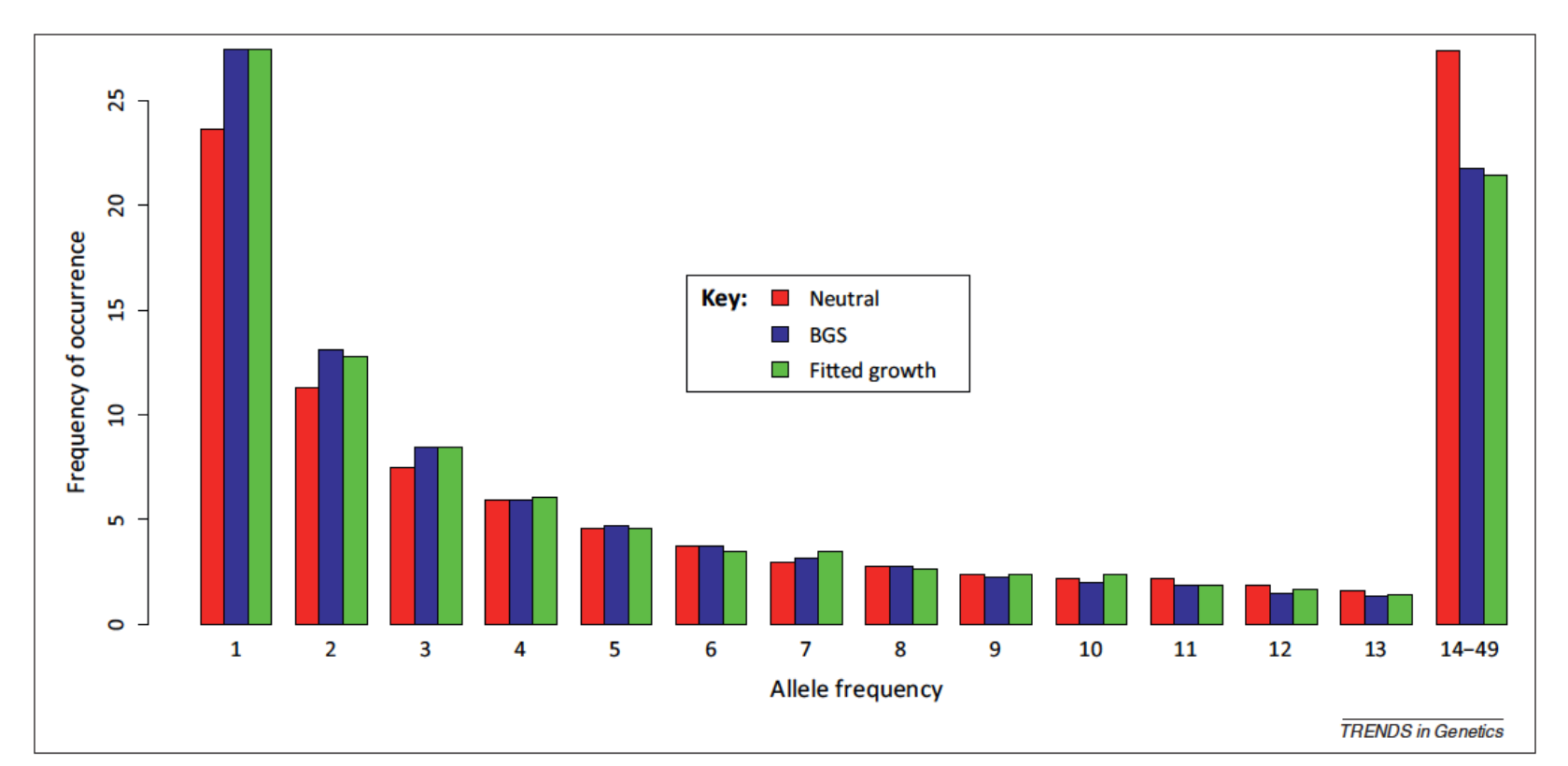
- thus, this process must be incorporated in to null expectations for patterns of variation along with demography



- Ewing & Jensen, 2016, *Mol. Ecol.*

So, to wrap up this section: Outstanding Issue #1 in population genetic inference: incorporating the effects of BGS

- **In other words, the same frequency spectra may be produced under models of BGS and, for example, neutral population growth**



- Bank et al. 2014, *TiG*

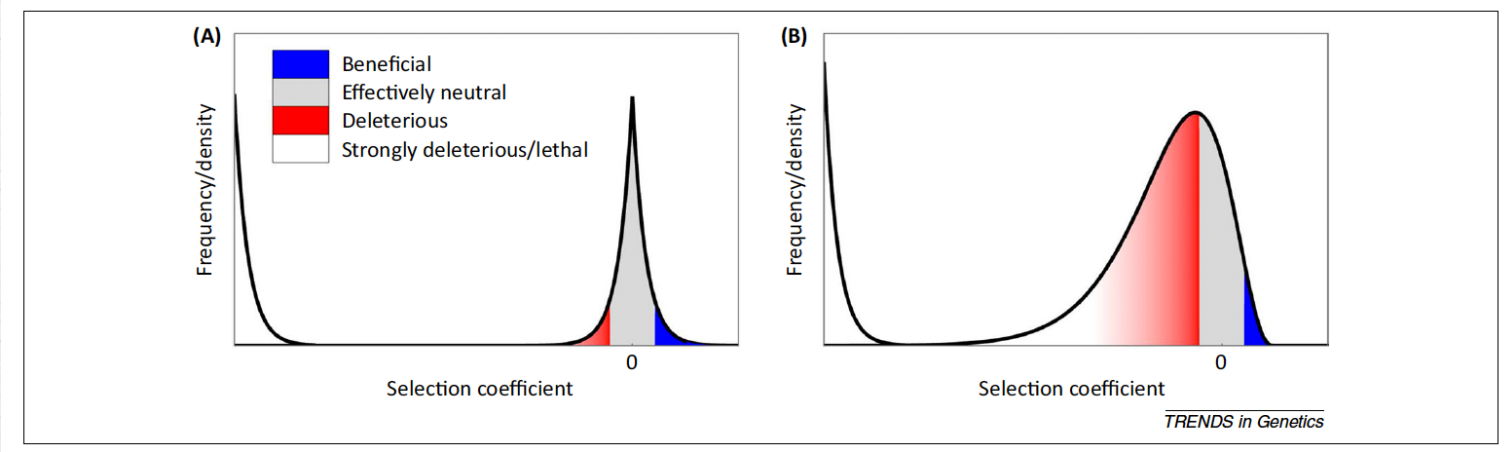
Outstanding Issue #2
in population genetic inference:
incorporating realistic DFEs

- **Incorporating realistic DFE's in to inference, in other words, better merging insights from experimental evolution in to empirical population genetics**

Outstanding Issue #2
in population genetic inference:
incorporating realistic DFEs

- **Incorporating realistic DFE's in to inference, in other words, better merging insights from experimental evolution in to empirical population genetics**

  - While it is common practice to assume that all mutations are of uniform *s*, or in the best case are given by an arbitrary distribution, we have the insights from experimental evolution to do much better than this



  - Bank et al. 2014, *TiG*

Outstanding Issue #2
in population genetic inference:
incorporating realistic DFEs

- **For example, the EMPIRIC system that we have developed over the last five years allows for the accurate experimental measurement of the full DFE for all possible new mutations**

- This work has also enabled us to understand how the DFE changes in the face of changing selective pressures on the population

C. Bank

Claudia will tell you more about this area later in the conference

- Hietpas et al. 2011, *PNAS*
- Hietpas et al. 2013, *Evolution*
- Bank et al. 2014, *Genetics*
- Bank et al. 2015, *MBE*

Outstanding Issue #3
in population genetic inference:
incorporating realistic offspring distributions

- **In a series of nice papers beginning with Eldon & Wakeley, the impact on neutral expectations of the site frequuency spectrum, linkage disequilibrium, and divergence have been explored for models with highly skewed offspring distributions**

Outstanding Issue #3
in population genetic inference:
incorporating realistic offspring distributions

- **In a series of nice papers beginning with Eldon & Wakeley, the impact on neutral expectations of the site frequuency spectrum, linkage disequilibrium, and divergence have been explored for models with highly skewed offspring distributions**

  - Atleast two things are clear:
  - 1) this skewness can strongly change neutral expectations, thus creating mis-inference when ignored
  - 2) such offspring distributions are biologically relevant for many organisms - ranging from viruses to plants to marine spawners

K. Irwin

see Kristen's poster for more details about this work

- Irwin et al., *Heredity*, in press

Thus, a revised best practices

- **Note that although we are not estimating BGS or MMC parameters yet (though work is underway), it is still possible to simulate them in order to address their impact on demographic and selection inference in a given population of interest**
  - **as well as to simulate user defined DFEs**

Table 2. A selection of commonly used simulation tools and included features[a]

Name	ms	msHot	fastsimcoal	msms	SFSCode	SLiM	SimuPop
Refs	[44]	[45]	[47]	[46]	[40]	[50]	[38]
Type of simulator	Coalescent	Coalescent	Coalescent	Coalescent	Forward	Forward	Forward
Recombination	Y	HotSpots	Y	Y	Y	Y	Y
Migration	Y	Y	Y	Y	Y	Y	Y
Hard sweeps	N	N	N	Y	Y	Y	N
Soft sweeps	N	N	N	Y	Y	Y	N
Linked selection	N	N	N	N	Y	Y	Y
DFE[b]	N	N	N	N	Y	Y	Y
Full chromosomes	N	N	Y	N	N[c]	Y	Y
Time samples	N	N	Y	Y	Y	Y	N

- Bank et al. 2014, *TiG*

Part II:
A more in depth example of the application of this work in our lab:
human cytomegalovirus (HCMV)

Cytomegalovirus exists across primates - from human to chimp to orang to macaque to African green monkeys

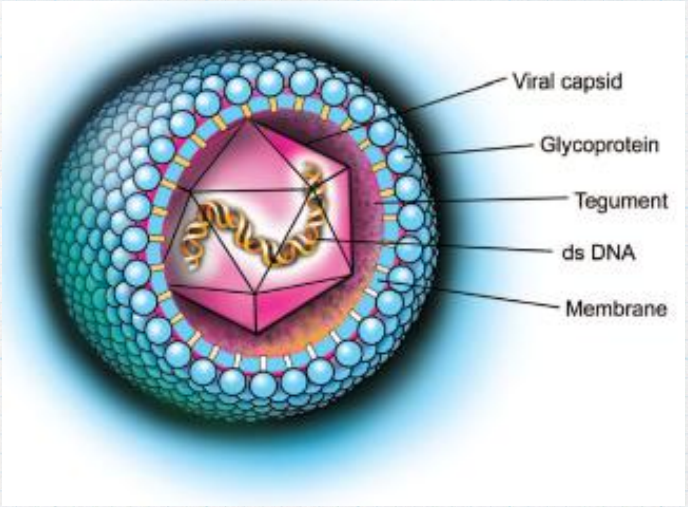
However, primate CMVs appear to be strongly species specific - with even chimpanzee (CCMV) and human (HCMV) viruses being unable to cross this recent species barrier

a bit about HCMV

- **a herpesvirus with seroprevalence of 30-90% of the global population**

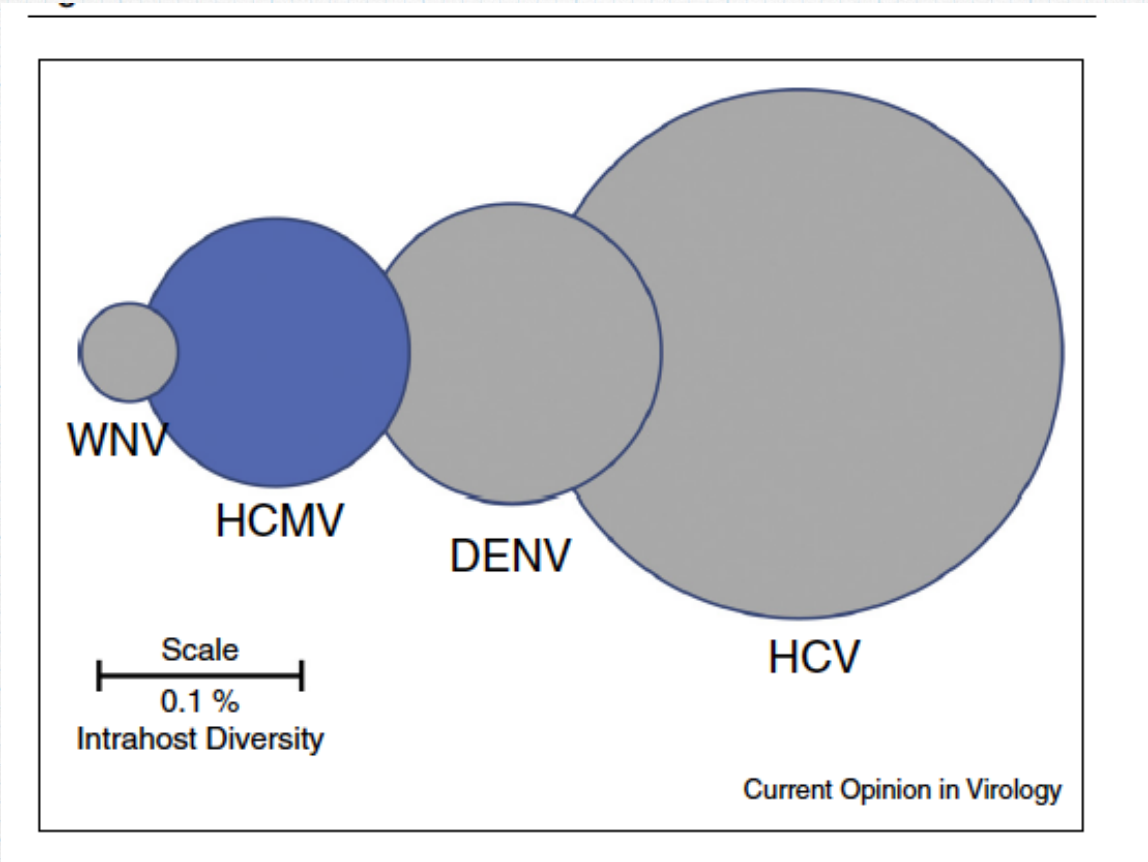
- 235kb DNA virus
- 200 open reading frames (perhaps as many as 700)

- primary infection usually via mucosal surfaces
  - after infection, remains latent within the body throughout life, and can be reactivated



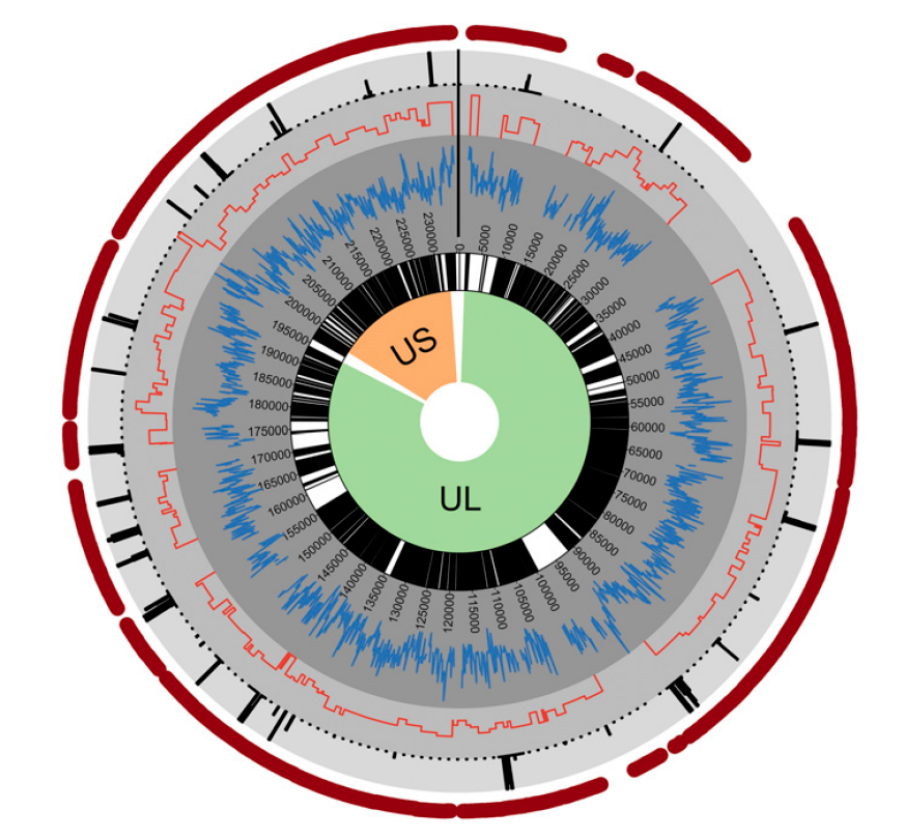
HCMV diversity compared to other viruses

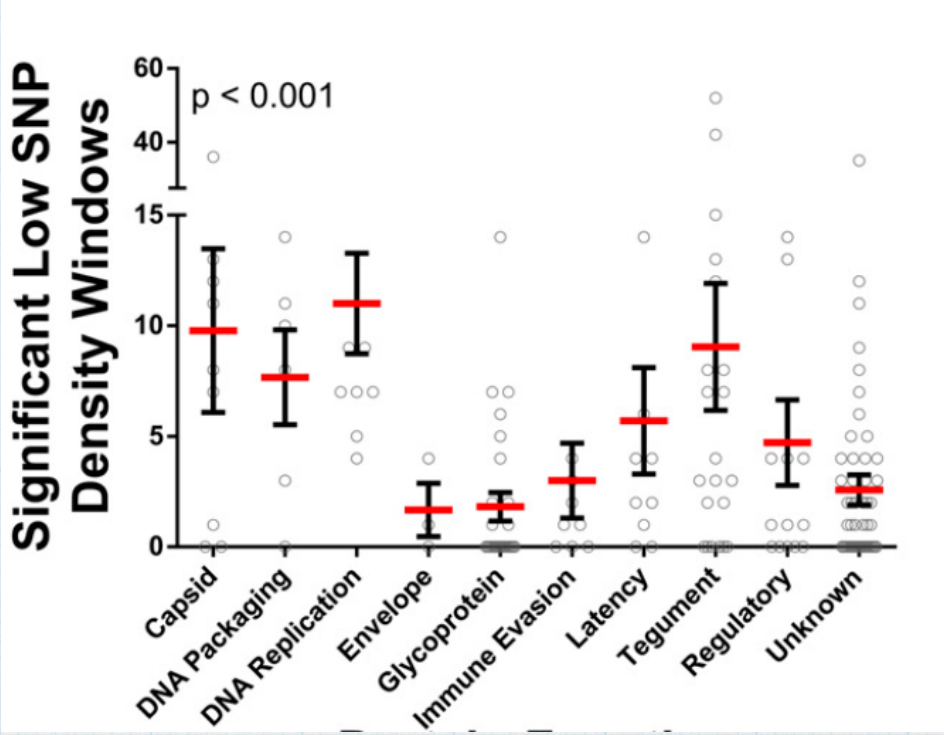
- **HCMV harbors high diversity for a DNA virus, on par in fact with RNA viruses**



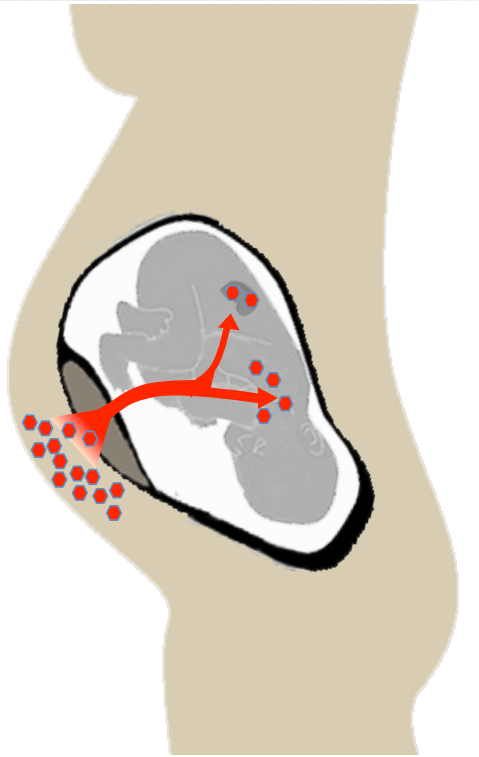
Diversity within HCMV

- We know quite a bit about how diversity varies across the genome, and the regions under strong functional constraint



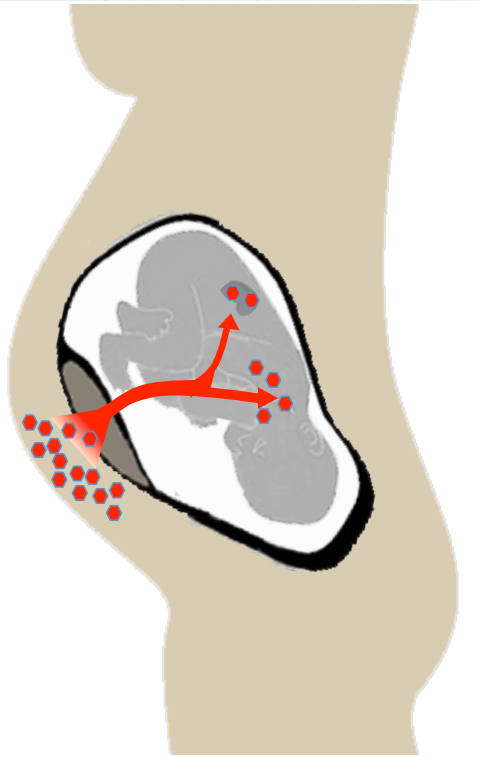


HCMV compartmentalization = within-host population structure



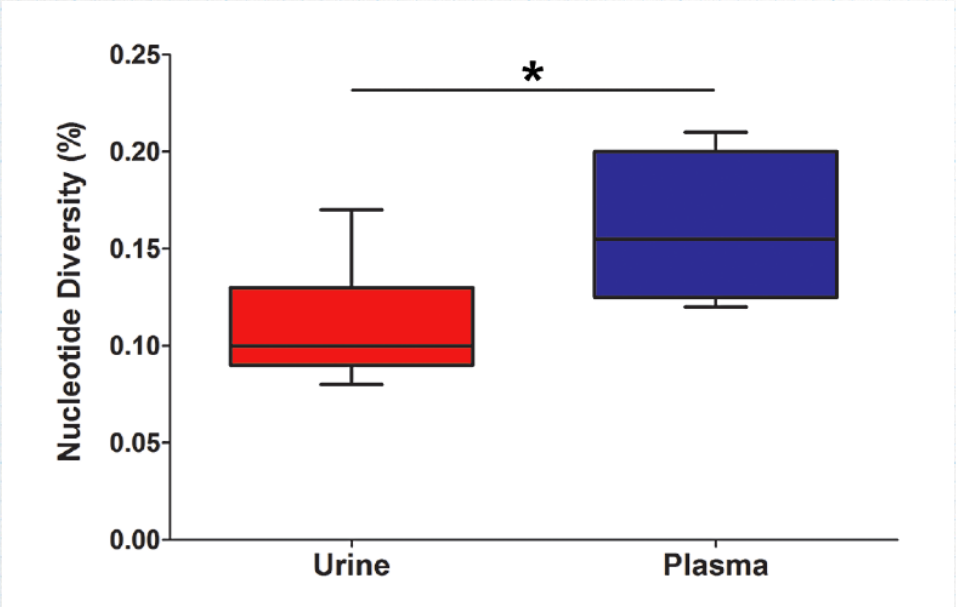
- **congenital infections are common, as the virus crosses the placenta and invades tissues throughout the fetus**
    - in fact, HCMV is the leading cause of infection-related birth defects

HCMV compartmentalization = within-host population structure



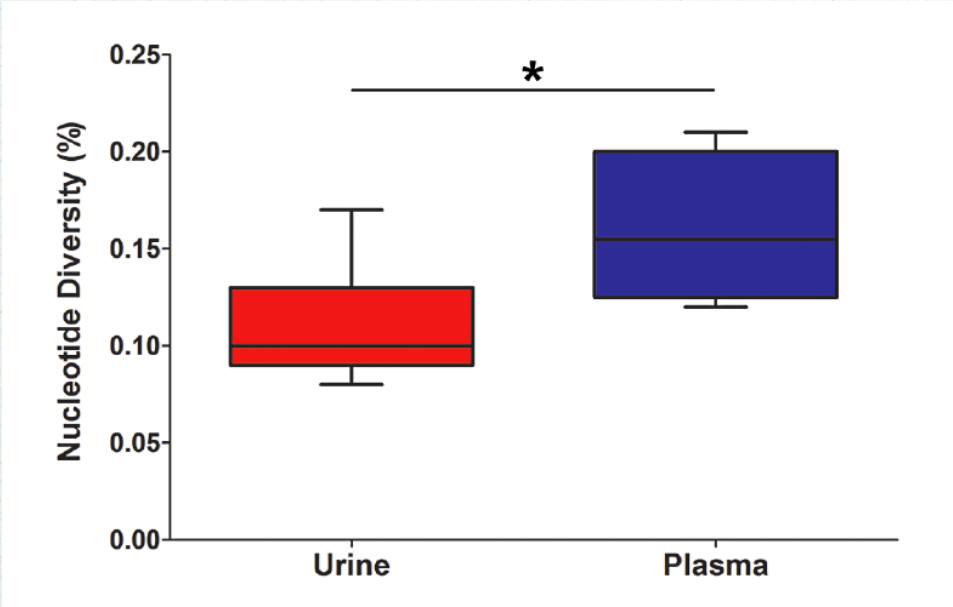
- **congenital infections are common, as the virus crosses the placenta and invades tissues throughout the fetus**
    - in fact, HCMV is the leading cause of infection-related birth defects
- we focus on congenital infections, in which the virus infects the fetus via the plasma, and then compartmentalizes - where these compartments can then be sampled post-birth from blood, urine (i.e., kidney compartment), and saliva (i.e., salivary gland compartment)
    - many other compartments exist, but are simply difficult (practically) to sample in infants

This population structure is indeed visible in genomic variation

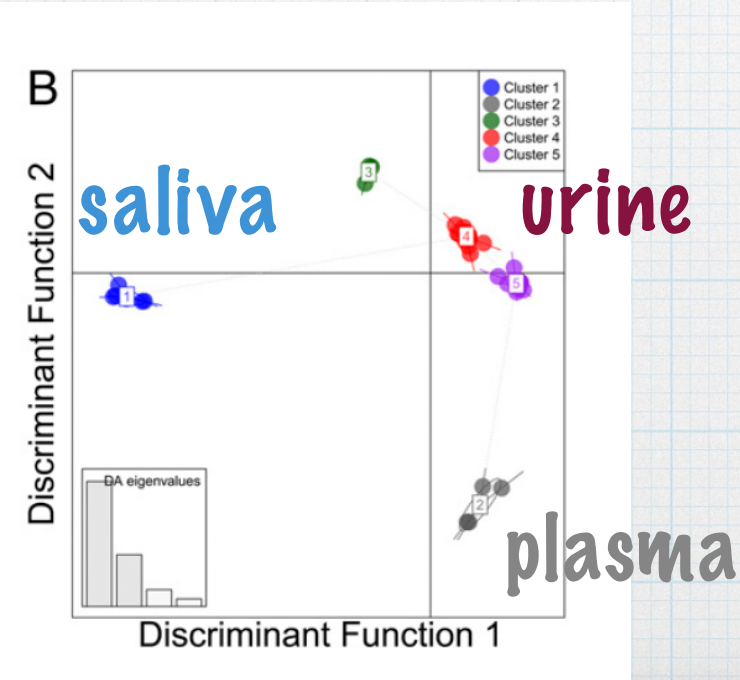


- Firstly, different compartments reliably harbor different levels of variation within patients

This population structure is indeed visible in genomic variation



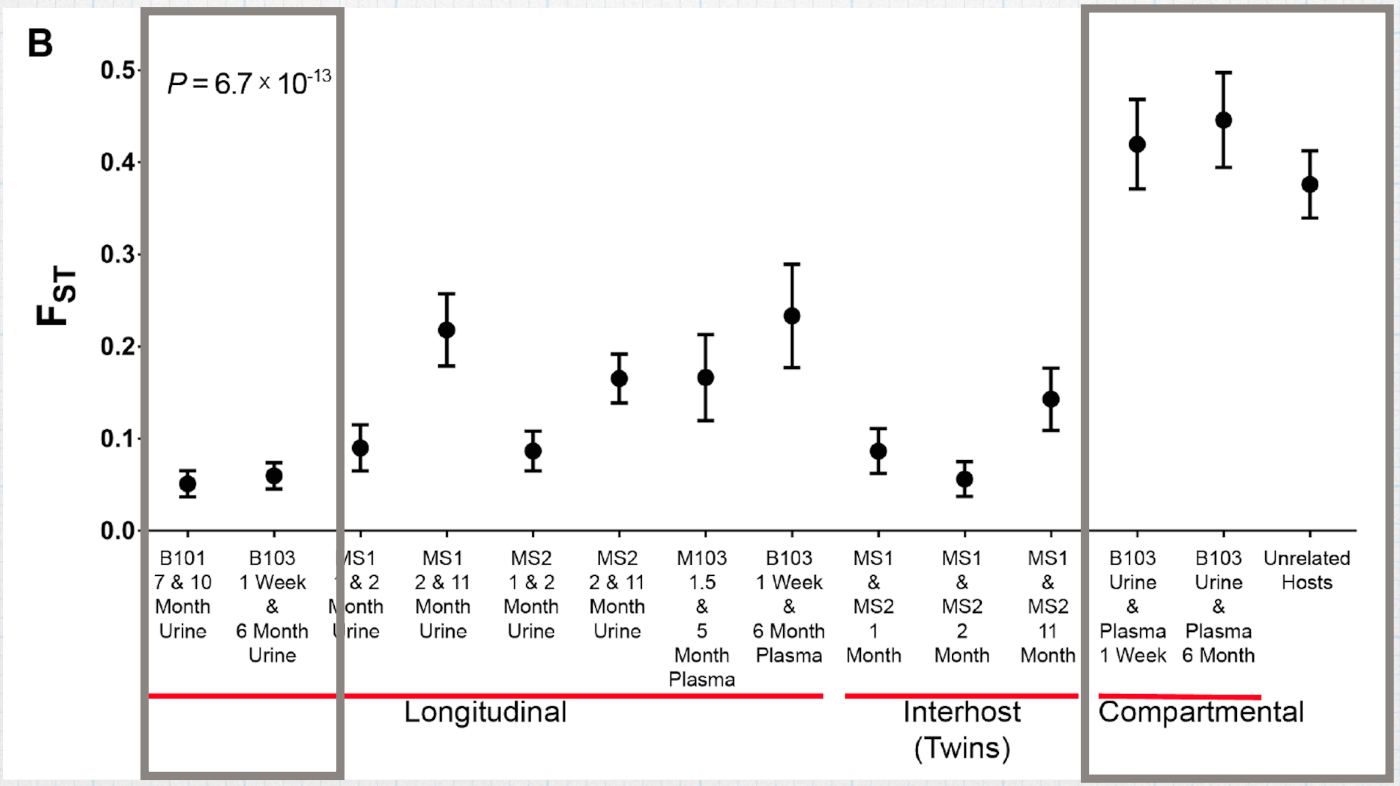
- Firstly, different compartments reliably harbor different levels of variation within patients



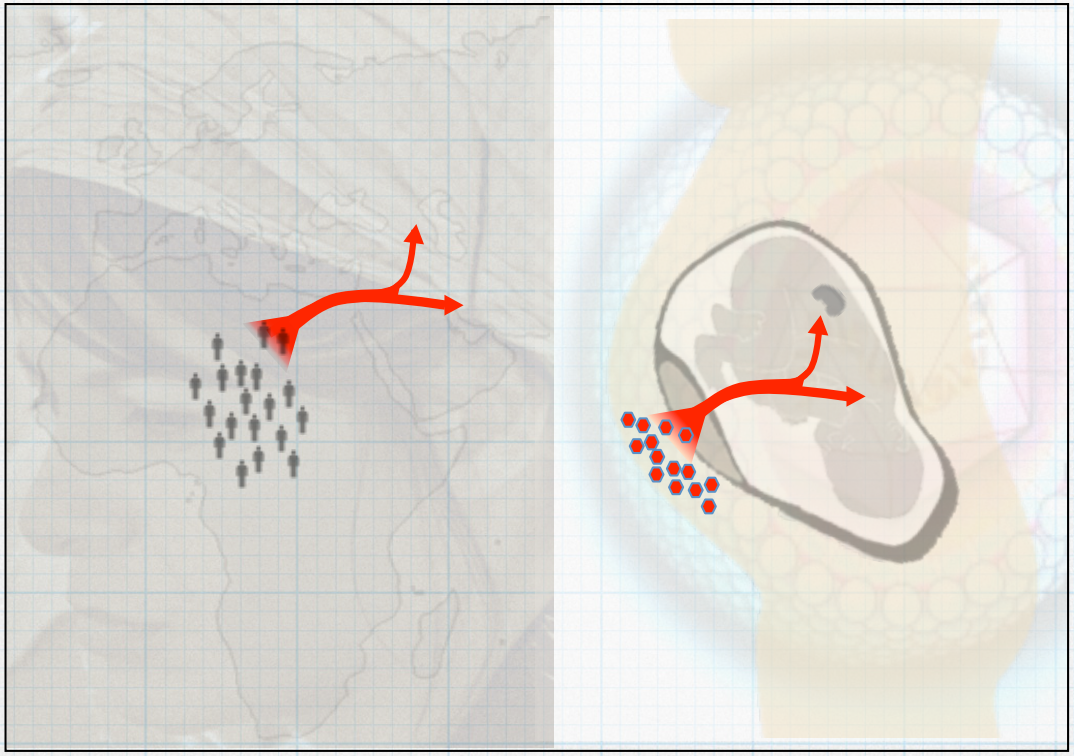
- Additionally, the compartments may be differentiated via PCA

These inter-compartmental differences are striking, resulting in highly differentiated populations within a single host

- **The differentiation between the urine and plasma compartment within a single host for example, is as different as between the urine and plasma compartments of unrelated individuals**



This has all been simply descriptive, so how about actually inferring the demography of infection?



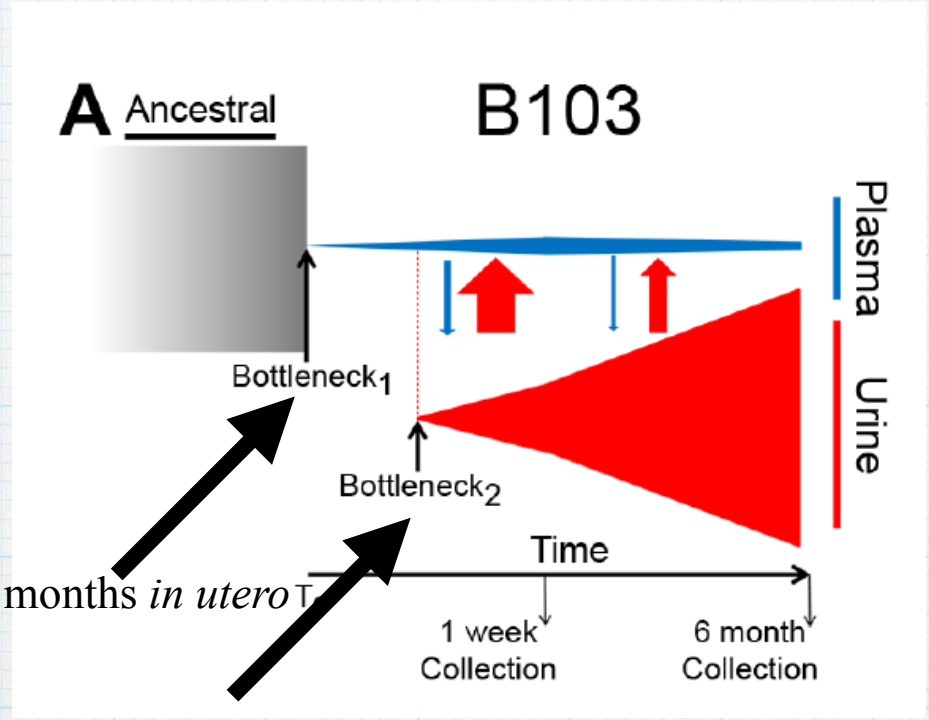
- **The demographic history of HCMV colonizing a novel host (e.g., fetus), is thus not entirely unlike the demographic history of humans colonizing a novel continent**
  - there is a larger ancestral population (mom / Africa)
  - there is a population size change associated with colonization (of mid-East / fetus)
  - these populations subsequently colonize further areas (globally / compartmentally)
  - colonized population may subsequently adapt to their new habitat
  - and migrants may be exchanged between these populations

On inferring the demography of infection

- With this type of data, in the same way that we estimate when humans colonized the Americas (and the associated bottleneck and migration rate), we can infer when HCMV colonized the fetus, when it subsequently colonized separate compartments within the fetus, and the bottlenecks and migration rates associated

- Thus, following the described inference approaches described above, we first worked to identify a well-fitting demographic model

Inferring population size change, the timing of population splits, and migration

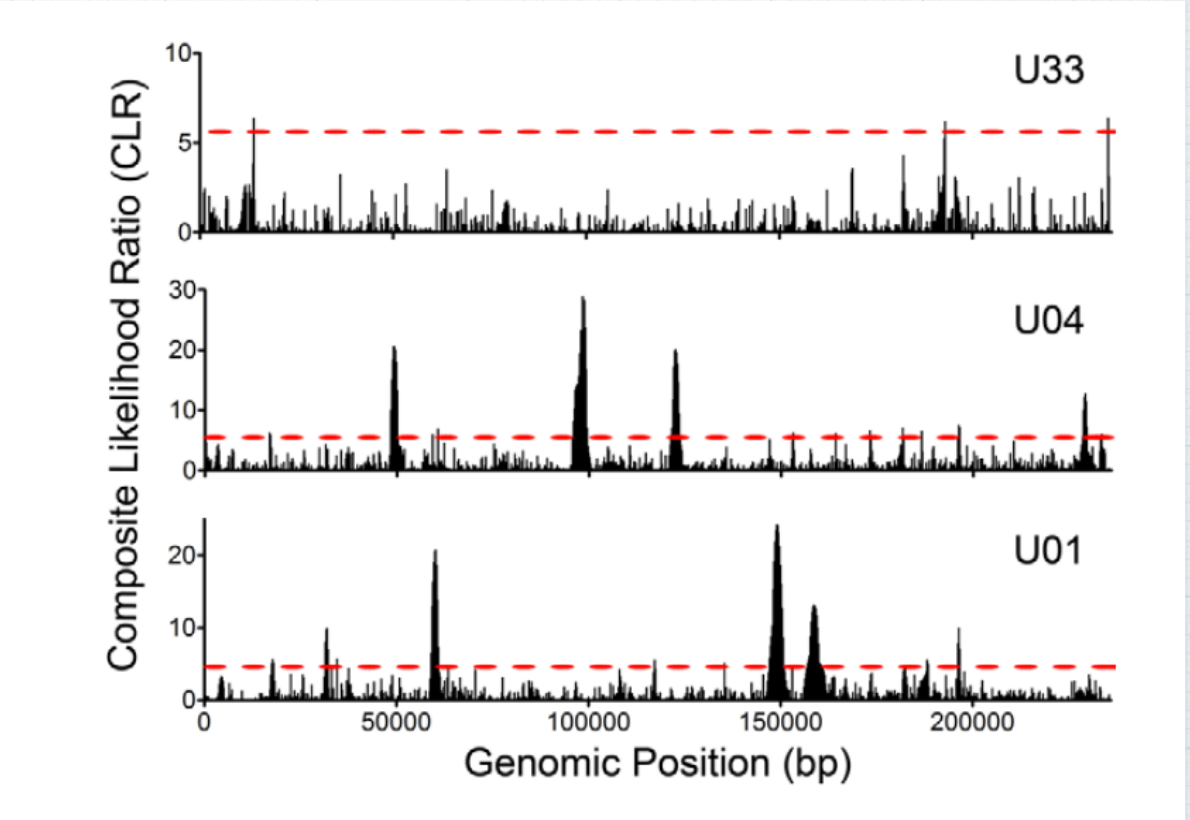


- **We consistently estimate a first bottleneck in to the plasma likely associated with initial infection, followed by subsequent bottlenecks during compartmentalization. We additionaly estimate compartment-specific effective population sizes and migration rates.**

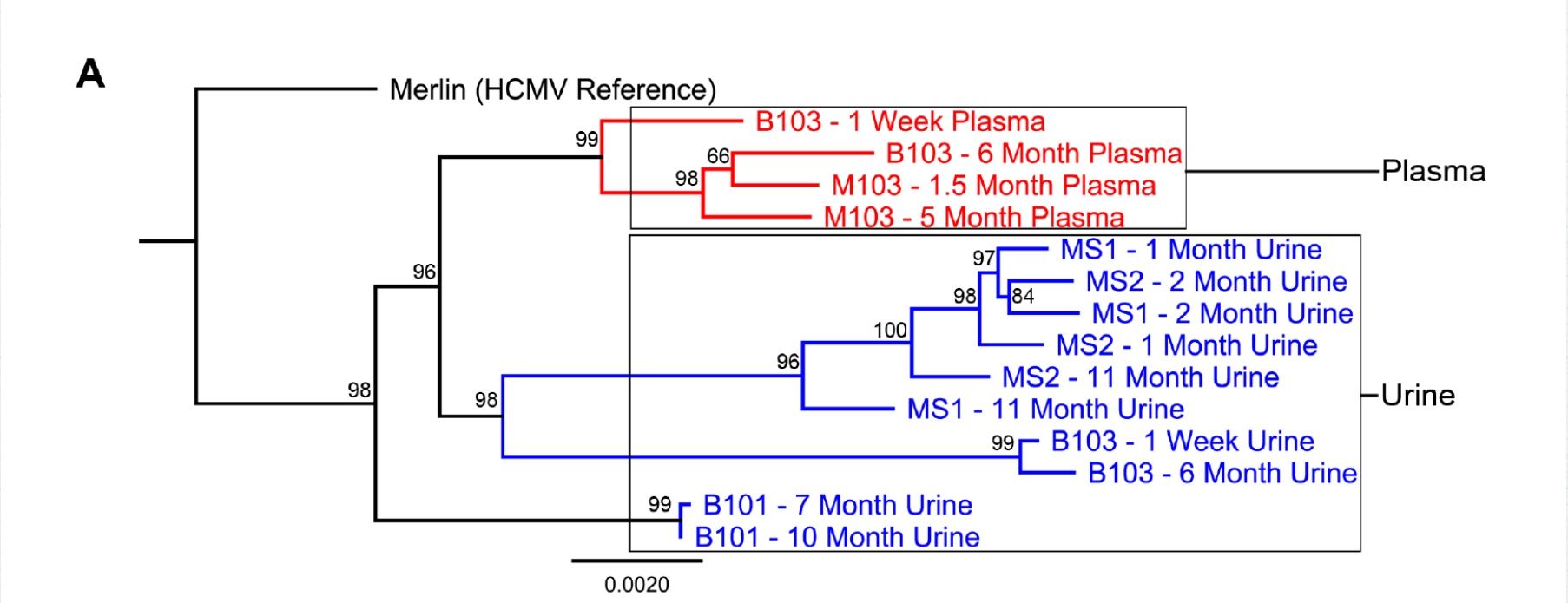
With this demographic model in hand, we can then evaluate our ability to detect hitchhiking patterns

- As described above, we first use simulations to determine our power to detect selection under the inferred demographic model, as well as the associated false-positive rate

- Taking multiple approaches (with the CLR shown here), we can catologue regions with hitchhiking patterns associated with infection and colonization



Moreover, we see strong compartment-specific convergence between patients



- in fact, the compartmental environment is sufficiently different, and associated selective pressure sufficiently strong, that the plasma populations (and urine populations) between patients group more closely than the urine and plasma population within a single patient (i.e., parallel adaptation)

We are also currently evaluating our ability to detect multiple infections

- **Current work is underway to evaluate our ability to detect mixed infections (i.e., essentially admixture mapping), that is - when multiple different strains are passed to the fetus**

- This work is particularly important clinically as, while it is unclear why some infected newborns are symptomatic and others are not, there is intriguing evidence suggesting that this may be related to multiple vs. single infections

In terms of efforts to understand the role of BGS in shaping patterns of variation, progress is well underway

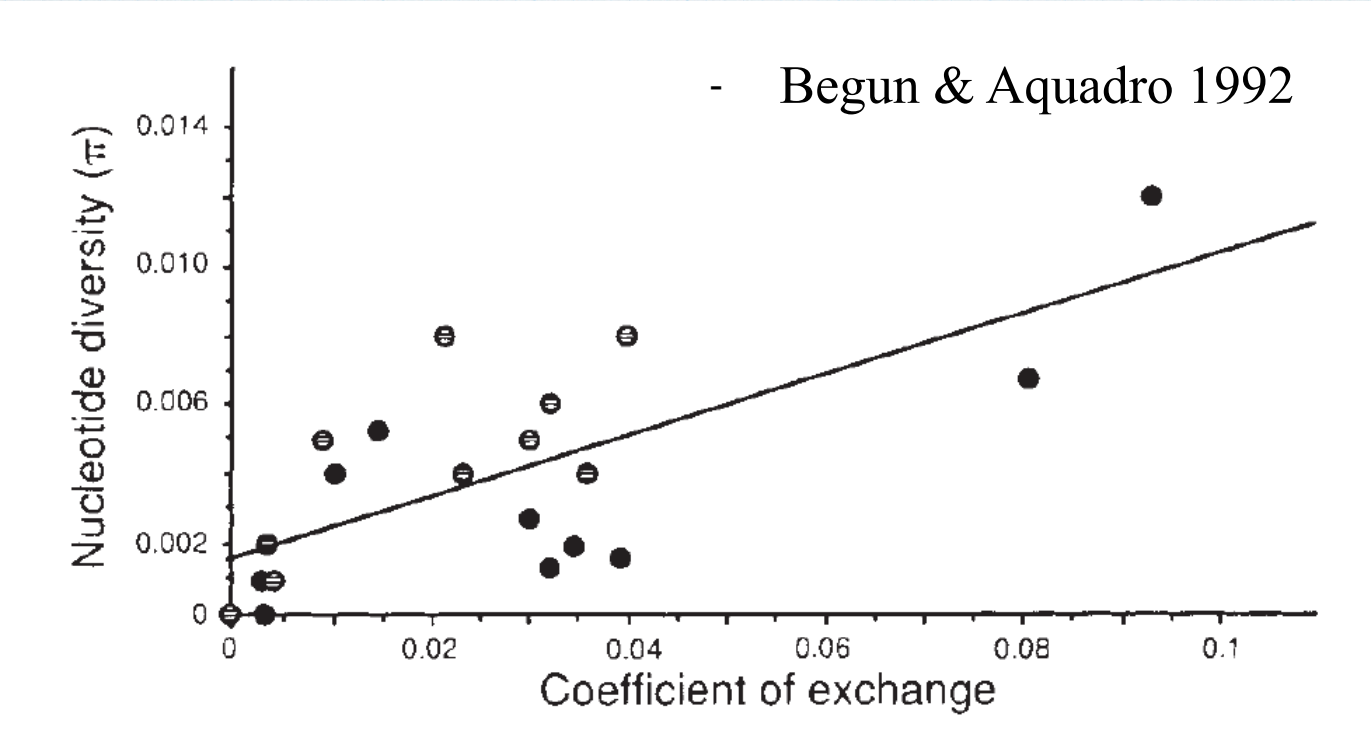
- One of the most universal patterns in population genetics is the relationship between diversity and recombination, and has been observed from Drosophila to *C. elegans* to humans

In terms of efforts to understand the role of BGS in shaping patterns of variation, progress is well underway

- One of the most universal patterns in population genetics is the relationship between diversity and recombination, and has been observed from Drosophila to *C. elegans* to humans

    - This observation indeed sparked the development of background selection by Charlesworth, as an alternative to genetic hitchhiking to explain the observation

        - Begun & Aquadro 1992

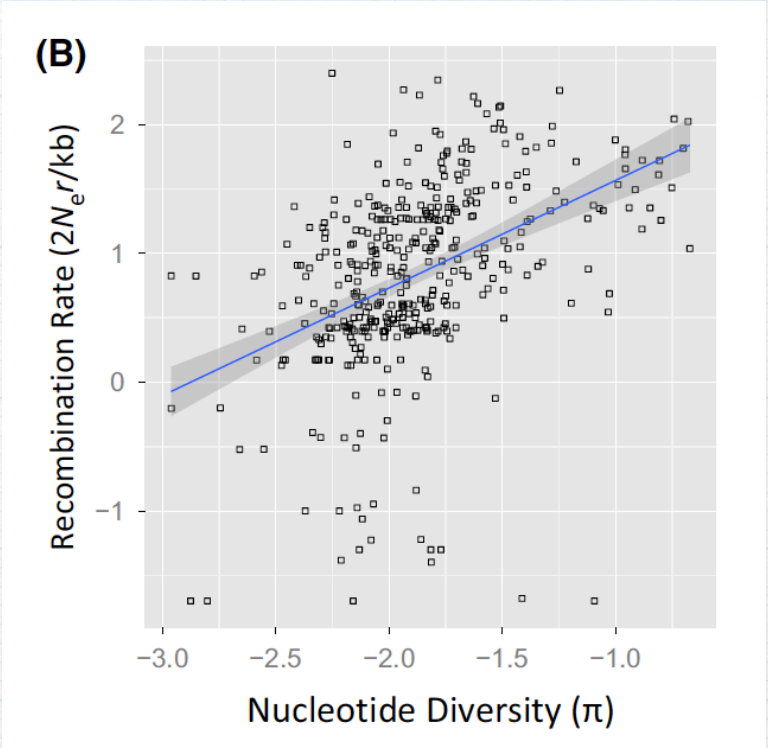


HCMV - a peek at the impact of BGS

- **We indeed observe this relationship in HCMV**

  - Utilizing prediction of Innan & Stephan, we find that this pattern is likely primarily driven by background selection

- Thus, along with the demographic estimates, this estimated rate of purifying selection inferred from these background selection patterns provides a much improved null model for hitchhiking mapping



To summarize, a few clinical implications of such population genetic analysis

- first clinically relevant estimate of the timing of fetal infection
  - (i.e., estimating the ages of viral colonization bottlenecks)
- first insights in to the genomic consequences of viral compartmentalization
  - (i.e., modeling population structure with migration and selection)
- first identification of infants multiply infected during pregnancy
  - (i.e., identifying viral population admixture events)

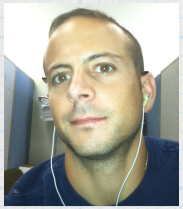
Renzette et al., The existence of the human cytomegalovirus quasispecies as revealed by high throughput sequencing. ***PLOS Pathogens*** (**2011**).

Renzette et al., Demography and selection contribute to the rapid evolution of cytomegalovirus within human hosts. ***PLOS Genetics*** (**2013**).

Renzette et al., Human cytomegalovirus intrahost evolution – a new avenue for understanding and controlling herpesvirus infections. ***Current Opinions in Virology*** (**2014**).

Renzette et al., On the limits and patterns of human cytomegalovirus genetic diversity in humans hosts. ***PNAS*** (**2015**).

Renzette et al., On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. ***Molecular Ecology*** (**2016**).



N. Renzette

Human cytomegalovirus (HCMV): a note on the next steps

- Given these results, future work in this area is focusing on the evaluation of experimental drug treatments which may have the capacity to prevent initial fetal infection (or at least greatly reduce the colonizing population size)

- All work to-date has considered congenital infections, but we are also evaluating horizontal infection, examining cohorts of infants infected in daycare and infections of immunocompromised individuals, in order to similarly characterize the demographic and selective history of horizontal infection

Human cytomegalovirus (HCMV): a note on the next steps

- Given these results, future work in this area is focusing on the evaluation of experimental drug treatments which may have the capacity to prevent initial fetal infection (or at least greatly reduce the colonizing population size)

- All work to-date has considered congenital infections, but we are also evaluating horizontal infection, examining cohorts of infants infected in daycare and infections of immunocompromised individuals, in order to similarly characterize the demographic and selective history of horizontal infection

Bottom line: this is not only a system where evolutionary analyses are providing highly valuable clinical information, but it is also a beautiful population genetic system for studying mutation/selection/drift/migration dynamics in a natural population

The Jensen Lab

Pre-doctoral lab members

CURRENT

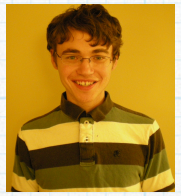

A. Kapopoulou


L. Ormond


H. Shim

FORMER


J. Crisci
postdoc
USC


A. Simkin
Asst Prof
Elon

Funding:

erc

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

DARPA

DEPARTMENT OF DEFENSE
UNITED STATES OF AMERICA

Post-doctoral lab members

CURRENT

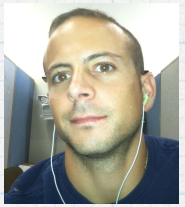

K. Irwin


S. Laurent


S. Matuszewski


V. Montano


S. Pfeifer


N. Renzette


S. Vuilleumier

FORMER


C. Bank
Group Leader
IGC


G. Ewing
Bioinformatician
Auckland


A. Ferrer
Bioinformatician
Barcelona


M. Foll
Group Leader
Lyon


L. Mathew
Bioinformatician
Tubingen


Y. Poh
Postdoc
MIT


C. Pokalyuk
Asst Prof
Magdeburg


D. Wegmann
Asst Prof
Fribourg

Hiring

The Jensen Lab will soon be advertising for multiple positions at both the postdoc and PhD student level, for 2017 start-dates

Arizona State University
School of Life Sciences
Center for Evolution & Medicine
http://jjensenlab.org