Linkage decay within a sequence similarity cloud of an oceanic *Vibrio* population

Matthew J. Melissa*

Department of Physics, Massachussetts Institute of Technology, Cambridge, Massachussetts 02139, USA

(Dated: May 13, 2016)

A major question at the intersection of microbial ecology and evolution asks to what extent recombination shapes the evolutionary dynamics of wild bacterial communities. In this article, we re-examine whole-genome sequence data collected from a population of oceanic *Vibrio cyclitrophicus.* We find supporting evidence that rampant recombination enables genes to sweep horizontally through populations. When we restrict our analysis to a main cloud consisting of only sufficiently similar sequences, we observe more rapid linkage decay, implying that recombination occurs preferentially among genetically similar individuals.

Keywords: genome-wide sweep, gene-specific sweep, recombination, linkage, sequence similarity cloud

Classic microbial evolution models often assume that selection dominates recombination, so that when a beneficial mutant fixes in a population, diversity along the entire genome is purged [1]. Alternatively, if recombination dominates selection, a beneficial mutation can recombine onto several distinct genetic backgrounds over the course of its fixation, preserving genetic diversity along the rest of the genome [2]. It remains unclear which factors determine whether genome-wide (diversity-purging) sweeps or gene-specific (diversity-preserving) sweeps will occur in wild microbial populations. A recent time-resolved metagenomic study identified observed a genome-wide sweep in a freshwater lake *Chlorobium* population, as well as several genetic signatures of prior gene-specific sweeps in other bacterial populations [3]. Genomic studies of ocean *Vibrionaceae* [7], as well as studies of *Synechococcus* [8] have also provided evidence of gene-specific sweeps.

Given empirically known recombination rates and selection coefficients, classic evolutionary models predict that gene-specific sweeps will not occur [2]. Two mechanisms have recently been proposed to explain the observation of gene-specific sweeps in wild populations: migration [4] and negative frequency-dependent selection [5]. Thus, evidence of gene-specific sweeps in a population implies that ecological mechanisms may play a significant role in the evolutionary dynamics.

In this paper, we re-examine ocean *Vibrionaceae* genomic data collected by Shapiro et. al. and analyzed in [7]. Applying statistical methods described in [8], we observe significant decay in genetic linkage with increasing SNP separation, providing supporting evidence for frequent recombination and the existence of gene-specific sweeps. In particular, we find evidence of a main "cloud" of similar sequences within which linkage decays considerably more rapidly.

THE DATA

Our analysis focuses on genetic diversity in a set of *Vibrionaceae* individuals studied in [7]. In [6], Hunt et. al. identified that within a population of *Vibrio*, preferential occurrence in large (L) vs. small (S) size-fractions of sea water is correlated with genetic content in a few regions of the genome. They hypothesized that the L and S strains occupy distinct ecological niches, with L strains preferentially attaching to large particles, and free-floating S strains more equipped to disperse over fast time scales. In [7], Shapiro et. al. sequenced whole-genomes for 20 *Vibrio cyclitrophicus* isolates–13 L and 7 S–and 2 reference *Vibrio splendidus* isolates. By identifying 725 *ecoSNPs*–sites in the genome at which one variant was present in all S strains and another variant was present in all L strains–clustered in a few regions of a genome otherwise suggesting phylogenetic intermingling of L and S strains, they inferred the prior occurrence gene-specific sweeps. In this project, our input data is the set of 22 aligned and assembled *Vibrio* genomes, partitioned into 954 locally collinear blocks (LCBs). (LCB blocks were chosen by Shapiro et. al. such that each LCB supports a particular phylogenetic relationship among the isolates.) We have calculated some basic diversity statistics [9] for this data, listed in Table I.

Genome length	3.5 Mbp
Number genes	4257
Number segregating sites S	128 Kbp
Avg. pairwise heterozygosity (per site)	0.0086
Number locally collinear blocks (LCBs)	954
Avg. LCB length	3.7 Kbp

TABLE I: Basic diversity statistics

SEQUENCE SIMILARITY CLOUD

In a recent publication [8], Rosen et. al. used metagenomic sequence data from a community of cyanobacteria (*Synechococcus sp.*) to infer the existence of a broad niche occupied by a sexual population. They find that when they restrict their analysis to a main "cloud" of the sequence data, linkage decays over considerably shorter lengths in genome. Rosen et. al. define their diversity "cloud" as follows: for each locus (approximately 500 bp) of their genome, a consensus strain is determined, consisting of the most common nucleotide at each of the base pairs in the locus. For each locus, all strains sufficiently similar to the consensus are marked as being within the "cloud"; others are excluded from the analysis. Thus, the main "cloud" may consist of a different number of strains at the various loci.

Here we seek to determine if a similar effect is present in *Vibrio* strains. Clearly the definition of a main "cloud" is rather ad hoc, and depends on a choice of genomic length scale for the different loci as well as a choice of sequence similarity threshold C below which strains are excluded from the cloud. (More precisely, a strain is excluded from the cloud if the fraction of nucleotides matching the consensus is less than C.) We define a "locus" to be a LCB on the basis that each LCB supports a unique phylogeny, meaning that the different LCBs can roughly be considered to be unlinked. We vary the sequence similarity threshold C from 0 (corresponding to all strains belonging to the cloud at each locus) to 0.99 (corresponding to sequences more than 1% diverged from the locus consensus being excluded from the cloud). In Figure 1, the number of strains excluded from the cloud is plotted for the various loci, at various sequence similarity thresholds C. We see that even for $C = 0.99$, a majority of strains are included in the cloud at 90% of loci.

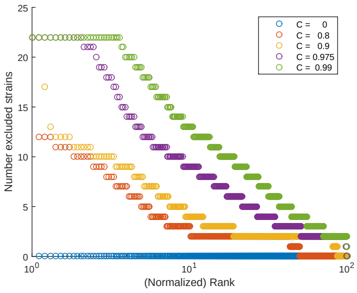

FIG. 1: Number of strains excluded from the main "cloud" at rank-ordered loci. As the sequence similarity threshold C is increased, a larger number of strains are excluded from the cloud. For $C \geq 0.8$ the number of strains excluded appears to decrease logarithmically with locus rank. Note that for the full set of data ($C = 0$) no strains are excluded.

LINKAGE STATISTICS

Here we present statistics describing the correlation of different sites in the genome, as a function of the SNP separation (measured in base pairs). We will focus all of our analysis on pairs of SNPs *within* loci (LCBs), since the analysis of linkage of SNPs on different LCBs has already been carried out in [7]. In addition, we find that linkage decay typically saturates on distance scales on the order of LCB length, suggesting that linkage of sites within LCBs contains the most interesting behavior. The primary statistic we will use to describe the extent of linkage is the coefficient of determination

$$< r^2 >=< \frac{(f_{ab} - f_a f_b)^2}{f_a(1-f_a)f_b(1-f_b)} >$$

where the average is carried out over pairs of polymorphic sites. For a particular pair of sites, f_a denotes the minor allele frequency of one of the sites, f_b denotes the minor allele frequency of the other site, and f_{ab} denotes the frequency of individuals with the minor allele at both sites. Clearly if all of the sites are evolving independently, then $< r^2 >= 0$, and if the genome is completely linked–that is, $f_a = f_{ab} = f_b$–then $< r^2 >= 1$. Restricting the average to polymorphic sites a distance d apart, we expect $< r^2 >$ to decrease with increasing d, since as the distance between polymorphic sites is increased, there is greater chance of a recombination event unlinking the two sites.

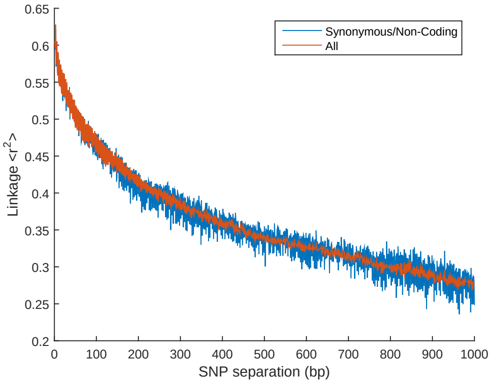

FIG. 2: Linkage correlation coefficient $< r^2 >$ for synonymous/non-coding SNP pairs and for full data set. As SNP separation is increased, linkage decays from a maximum value $< r^2 > \approx 0.6$. The linkage decay pattern for synonymous/non-coding mutations appears to match that of the entire data set, although increased variance is observed for the synonymous/non-coding mutations. Note that the pattern exhibits oscillations, measured to have a period of 3 bp, which can be attributed to non-random codon frequencies across the genome.

In Figure 2, we plot $< r^2 >$ as a function of SNP separation and observe the characteristic decay of linkage as a function of SNP separation. Following the approach in [8], we plot $< r^2 >$ averaged over pairs of synonymous and non-coding mutations (identified using gene annotations available at NCBI GenBank), as well as $< r^2 >$ averaged over all pairs of mutations. (Because of time limitations, we considered all 30,035 non-coding mutations but only the subset of 31,539 synonymous mutations located on genes coded in one particular direction and confined to a single LCB. A future study could easily consider all synonymous mutations but the results should be similar.) Restricting the analysis to pairs of synonymous/non-coding mutations eliminates the possibility that observed linkage could be an artifact of epistatic interactions among mutations. However, we see that the linkage decay behavior for synonymous/non-coding mutations is largely indistinguishable from the behavior for all mutations (apart from increased variance, which is presumably because there are fewer synonymous/non-coding mutations), so for the remainder of this paper we will consider linkage among all pairs of mutations.

Figure 3 displays the linkage $< r^2 >$ decay for the main "clouds" corresponding to various sequence similarity thresholds C. The relationship between C and linkage is not necessarily monotonic, but the cloud corresponding to $C = 0.99$ is significantly less linked than the full data set.

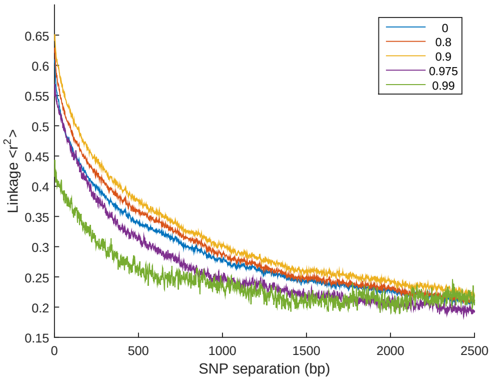

FIG. 3: Decay of linkage correlation coefficient r^2 as a function of SNP separation. The legend displays the various sequence similarity thresholds. Note that while the most exclusive similarity thresholds $C = 0.975$ and $C = 0.99$ correspond to the least linkage, the relationship between linkage decay and sequence similarity threshold is in general non-monotonic. We also see that linkage saturates at around $\delta \approx 2000$ bp, suggesting a typical length scale for recombined genomic fragments.

While Figure 3 considers linkage averaged over pairs of mutations scattered throughout the genome, we can also consider the variation in linkage decay lengths throughout the genome. We define the linkage decay length L_D for a particular locus (in our case, LCBs) as follows: L_D is the minimum distance l such that median$(r^2) < e^{-1}$, where the median is taken over all SNP pairs (on the locus) closer than l bp apart. Note that we are taking of the median of r^2 for individual SNP pairs, not of r^2 averaged over distances. In addition, we only consider pairs of SNPs on loci with more than 50000 SNP pairs. A rank-ordered plot of the linkage decay length L_D at the different loci is shown in Figure 4, for a range of sequence similarity thresholds. A few features of the plot are noteworthy: The linkage decay length varies over four orders of magnitude, with the maximal decay length roughly independent of choice of sequence similarity threshold C. A significant fraction of loci are effectively unlinked, with a decay length $L_D = 1$. As C is increased above 0.9, the typical L_D across the genome drops off very rapidly.

Finally, we can examine the full joint distribution of minor allele frequencies f_a and f_b, using the method described in [8]. In particular, we calculate the SNP pair enrichment E_{f_a,f_b} for bins corresponding to minor frequencies f_a and f_b, as follows: Let N_{f_a,f_b} be the number of SNP pairs such that on one site, the minor allele is present at a frequency f_a, and on the other site, the minor allele is present at a frequency f_b. Let n_{f_a} be the number of SNP sites at which the minor allele is present at frequency f_a, and define n_{f_b} analogously. Then define

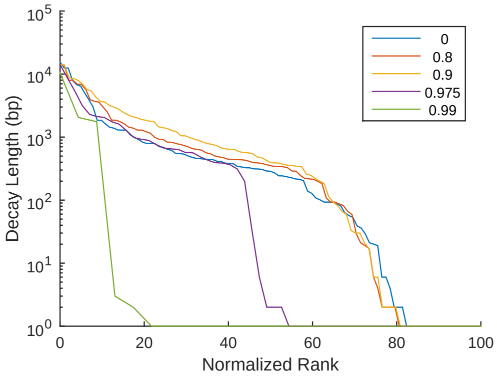

FIG. 4: Linkage decay length L_D at rank-ordered loci. Here we see evidence of significant variation in the linkage decay length throughout the genome. The legend displays cloud sequence similarity thresholds C ranging from 0 (full data set) to 0.99. For $C = 0.975$ and $C = 0.99$, a large fraction of loci are effectively unlinked. (For example, for $C = 0.99$, $L_D = 1$ bp at almost 80% of loci.)

$$E_{f_a,f_b} = \log \frac{2N_{f_a,f_b}}{n_{f_a} n_{f_b}}$$

That is, E_{f_a,f_b} denotes the logarithmic ratio of the number of SNP pairs with minor allele frequencies f_a and f_b to what would be predicted from the site-frequency spectrum. The resulting colored histogram is shown in Figure 5, for both the full set of SNP pairs (top) and the main cloud of SNPs corresponding to $C = 0.99$ (bottom). The most obvious feature is the predominance of large E_{f_a,f_b} values along the diagonal, which is consistent with (but not necessarily evidence of) strong linkage. We also see that the magnitude of this effect decreases with increasing SNP pair separation, and is almost entirely suppressed for SNP separation larger than 1000-10000 bp.

There are a few visible differences between plots for the full set of SNP pairs and the $C = 0.99$ cloud. First, SNP pair enrichment along the main diagonal is stronger for the $C = 0.99$ cloud than for the full set. This is surprising, since large E_{f_a,f_b} values are consistent with strong linkage, and Figure 3 clearly shows weaker linkage for the $C = 0.99$ cloud. It is important to note that for SNP pairs along the main diagonal, $f_a = f_b$, but this does *not* imply perfect linkage ($f_a = f_{ab} = f_b$, or equivalently, $r^2 = 1$). Second, we note that the frequency histogram for the full set has a block-diagonal form, while the in the $C = 0.99$ case, enrichment rapidly decays off of the main diagonal. This observation can be explained in the following way: Strains excluded from the main cloud likely obtained recent mutations, and these mutations perturb minor allele frequencies from equality. Future studies could predict the general features of these joint frequency histograms depending on relevant model parameters such as population size, mutation rate, and recombination rate.

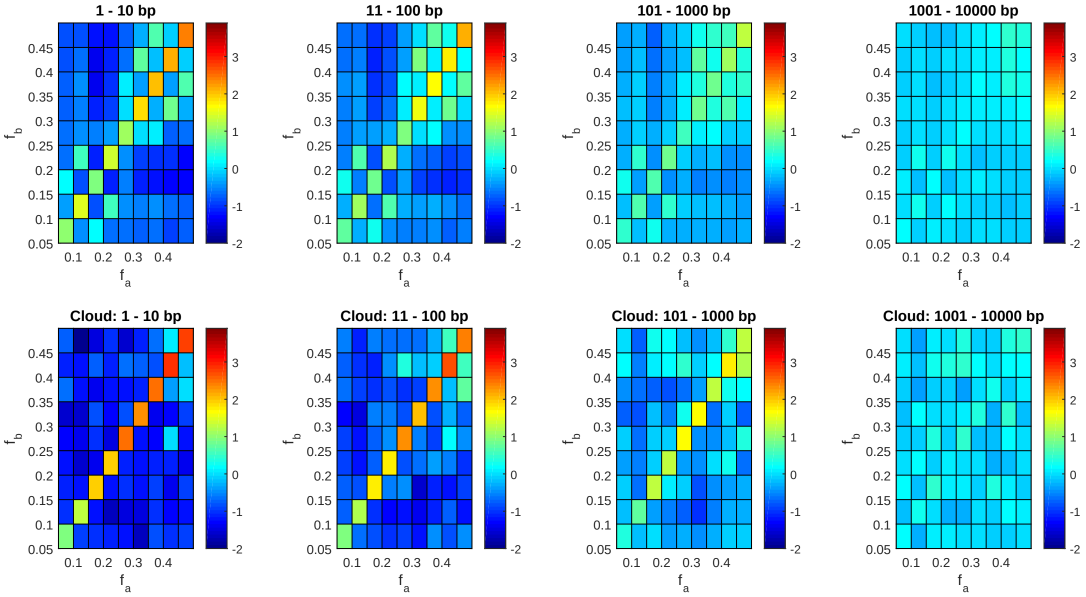

FIG. 5: Enrichment of minor allele frequency pairs, for various SNP pair separations. The top row displays frequency histograms for the full data set, and the bottom row displays frequency histograms for the $C = 0.99$ cloud. For a bin corresponding to frequencies f_a and f_b, dark red denotes that a greater number of SNP pairs are observed with minor allele frequencies f_a and f_b than would be expected by chance, given the single-site frequency spectrum (distribution of f_a).

DISCUSSION

Our analysis provides supporting evidence that this particular population of *Vibrio* is highly sexual, with recombination unlinking genes from their initial genetic backgrounds and enabling them to sweep through the population. In particular, we observe that when our analysis is restricted to a main "cloud" of genetically similar strains, linkage decay length falls off more rapidly with SNP separation. We can interpret this decreased linkage within the main cloud as evidence that recombination occurs more frequently among genetically similar individuals. Mathematical modelling has demonstrated that differential recombination rates–i.e. recombination rates that fall off with sequence divergence–can result in spontaneous formation of genetic clusters, thus playing a key role in ecological divergence [10].

Our results are not as quantitatively striking as the results for *Synechococcus* in [8], but further assessment of different locus length scales and sequence similarity thresholds may reveal a stronger effect. A future theoretical study could examine the relationship between a particular functional dependence of the recombination rate on sequence divergence, and linkage decay statistics within a well-defined sequence cloud.

There are a few limitations to the applicability of our results that are worth noting. First, our results are generally visibly apparent from the data, but rigorous statistical hypothesis testing is necessary to establish confidence that these features cannot be explained by chance fluctuations of an appropriate null model. In [8], Rosen et. al. present joint frequency histograms produced by an asexual drift simulation; we assume that an asexual drift simulation for the *Vibrio* sample would produce a similar result–that is, we expect the asexual drift simulation would look roughly like the 1001-10000 bp case of Figure 5. While the results in Figure 5 are visually interesting, further study is necessary to determine whether a joint frequency histogram is really appropriate for a sample size this small.

Second, we should note that each SNP pair in our analysis consists of two SNPs within the same LCB. This choice was made for computational reasons, and because analysis of SNP pairs extending across LCBs was conducted in [7]. It should be noted that LCBs were chosen by Shapiro et. al. on the basis that each LCB supports a unique phylogeny. That is, division of a region into two LCBs suggests that the two LCBs were at some point broken up by a recombination event that gave rise to distinct phylogenetic trees for the two LCBs. Importantly, this suggests that individual LCBs were *not* broken up by a recombination event that would suggest more than one phylogeny, and so we expect that linkage statistics for SNP pairs within LCBs may be significantly different than the statistics for the full set of SNP pairs. Interestingly, we still observed rapid linkage decay for SNP pairs within LCBs, implying that this bias may not be particularly serious. Future studies could determine whether our results apply to SNP pairs across the entire genome, or whether they are only applicable to SNP pairs on regions supporting a particular phylogeny.

In conclusion, our analysis provides preliminary evidence supporting the following picture for the *Vibrio* population studied: Rampant recombination enables genes to sweep through populations while preserving genetic diversity. Recombination occurs more frequently within a main "cloud" of similar sequences, potentially playing a key role in formation of genetic clusters and the first steps of ecological differentiation.

ACKNOWLEDGEMENTS

I would like to thank Daniel P. Rice and Michael Desai for useful discussions.

Assembled and aligned genomes, divided into locally collinear blocks (LCBs), were obtained from http://almlab.mit.edu/starrinights.html. Gene annotations were obtained from NCBI GenBank, BioProject PRJNA81863.

[1] F. M. Cohan, "Bacterial Species and Speciation," *Systematic Biology* **50**, 513 (2001).

[2] B. J. Shapiro, L. A. David, J. Friedman, and E. J. Alm, "Looking for Darwin's footprints in the microbial world," *Trends in Microbiology* **17**, 196 (2009).

[3] M. L. Bendall, S. LR Stevens, *et. al.* "Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations," *The ISME Journal*, Advance online publication (2016).

[4] R. Niehus, *et. al.* "Migration and horizontal gene transfer divide microbial genomes into multiple niches," *Natu. Commun.*, **6:8924** (2015).

[5] N. Takeuchi, O. Cordero, E. V. Koonin, and K. Kaneko "Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection," *BMC Biology*, **13** (2015).

[6] D. A. Hunt, L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz, "Resource Partitioning and Sympatric Differential Among Closely Related Bacterioplankton," *Science* **320**, 1081 (2008).

[7] B. J. Shapiro, J. Friedman, O. Cordero, S. Preheim, S. Timberlake, G. Szabo, M. Polz, and E. Alm, "Population Genomics of Early Events in the Ecological Differentiation of Bacteria," *Science* **336**, 48 (2008).

[8] M. J. Rosen, M. Davison, D. Bhaya, and D. Fisher, "Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche," *Science* **348**, 1019 (2015).

[9] J. Wakeley, *Coalescent Theory*, Roberts and Company Publishers Greenwood Village, CO 13-18 (2009).

[10] W. P. Hanage, B. G. Spratt, K. M.E Turner, and C. Fraser, "Modelling Bacterial Speciation," *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 2039 (2006).

[*] Electronic address: mjmel@mit.edu