ON THE NEUTRALOME OF GREAT APES

AND NEAREST NEIGHBOR SEARCH IN METRIC SPACES

by

August Woerner

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF GENETICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by **August Woerner**, **ON THE NEUTRALOME OF GREAT APES AND NEAREST NEIGHBOR SEARCH IN METRIC SPACES** and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

John Kececioglu _____ Date: 07/28/2016

Michael Hammer _____ Date: 07/28/2016

Joseph Watkins _____ Date: 07/28/2016

Ryan Gutenkunst _____ Date: 07/28/2016

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 07/28/2016
Dissertation Director: Michael Hammer

_____ Date: 07/28/2016
Dissertation Director: John Kececioglu

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: August Woerner

ACKNOWLEDGEMENT

My deepest thanks goes to the many people that have supported me during my tenure at the University of Arizona. First I would like to thank my advisors, John Kececioglu and Michael Hammer. John, I will always admire your singular focus and precision. Michael, I will always admire your academic breadth and depth. I thank you both for your patience with me as a student. I joined Michael's lab in 2004, and therein I enjoyed 11 years of fun and excitement doing research in computation and population genomics. This 11 years was spent without having to worry about the finances and headaches that more senior researchers must battle with daily, and it was an incredible gift for which I am most thankful. John, together we have been researchers since 2009, and your guidance on how to think computationally will help me throughout my academic career. I entered the UA knowing little of the computation and genomics, and now that I leave I feel that while I still have much to understand, I now have the tools to learn what I need to succeed.

I would also like to thank my committee members. Joe, we have worked together for years, and your insights into applications of probability and statistics in genomics and population genetics continue to amaze me. Ryan, we have not collaborated much, but I deeply appreciate your guidance and sharp questions.

Members of the Hammer lab, past and present, have probably had a greater role in my academic development than have my committee. Members of the lab, such as Fernando Mendez, Krishna Veeramah, Tanya Karafet, Olga Savina, Murray Cox, Tesa Severson, Maya Pilkington, PingHsun (Benson) Hsieh, Ariella Gladstein, Consuelo Quinto are but a few of whom have fundamentally shaped my thinking both in the roles of student and teacher. I also thank the

Genetics department, and Cora Varas-Nelson in particular, for helping to make this degree possible.

Saving the best for last, I would like to thank my family. My sister, Margie, and my parents, Betty and John, thank you for all of your help throughout the years. My wife, Kelly, you are beautiful, amazing, strong and kind. My eldest, Micah, your heart is enormous and your love of your brother is awesome. And my youngest, Kai, you are (for a toddler) patient, joyful, loving and kind. I love you all very much, and if it weren't for your support this dissertation would not be possible.

Table of Contents

ABSTRACT

Problems of population genetics are magnified by problems of big data. My dissertation spans the disciplines of computer science and population genetics, leveraging computational approaches to biological problems to address issues in genomics research. In this dissertation I develop more efficient metric search algorithms. I also show that vast majority of the genomes of great apes are impacted by the forces of natural selection. Finally, I introduce a heuristic to identify neutralomes—regions that are evolving with minimal selective pressures—and use these neutralomes for inferences on effective population size and breeding sex ratios in great apes.

We begin with a formal and far-reaching problem that impacts a broad array of disciplines including biology and computer science; the *k*-nearest neighbors problem in generalized metric spaces. The *k*-nearest neighbors (*k*-NN) problem is deceptively simple. The problem is as follows: given a query q and dataset D of size n, find the k-closest points to q. This problem can be easily solved by algorithms that compute *k*th order statistics in O(n) time and space. It follows that if D can be ordered, then it is perhaps possible to solve *k*-NN queries in sublinear time. While this is not possible for an arbitrary distance function on the points in D, I show that if the points are constrained by the triangle inequality (such as with metric spaces), then the dataset can be properly organized into a dispersion tree (**Appendix C**). Dispersion trees are a hierarchical data structure that is built around a large dispersed set of points. Dispersion trees have construction times that are sub-quadratic ($O(n^{1.5} \log n)$) and use $O(n)$ space, and they use a provably optimal search strategy that minimizes the number of times the distance function is invoked. While all metric data structures have worst-case $O(n)$ search times, dispersion trees have average-case search times that are substantially faster than a large sampling

of comparable data structures in the vast majority of spaces sampled. Exceptions to this include extremely high dimensional space ($d > 20$) which devolve into near-linear scans of the dataset, and unstructured low-dimensional ($d < 6$) Euclidean spaces. Dispersion trees have empirical search times that appear to scale as $O(n^c)$ for $0 < c < 1$. As solutions to the *k*-NN problem are in general too slow to be used effectively in the arena of big data in genomics, it is my hope that dispersion trees may help lift this barrier. With source-code that is freely available for academic use, dispersion trees may be useful for nearest neighbor classification problems in machine learning, fast read-mapping against a reference genome, and as a general computational tool for problems such clustering.

Next, I turn to problems in population genomics. Genomic patterns of diversity are a complex function of the interplay between demographics, natural selection and mechanistic forces. A central tenet of population genetics is the neutral theory of molecular evolution which states the vast majority of changes at the molecular level are (relatively) selectively neutral; that is, they do not effect fitness. A corollary of the neutral theory is that the frequency of most alleles in populations are dictated by neutral processes and not selective processes. The forces of natural selection impact not just the site of selection, but linked neutral sites as well. I proposed an empirical assessment of the extents of linked selection in the human genome (**Appendix A**). Recombination decouples sites of selection from the genomic background, thus it serves to mitigate the effects of linked selection. I use two metrics on recombination, both the minimum genetic distance to genes and local rates of recombination, to parse the effects of linked selection into selection from genic and nongenic sources in the human genome. My empirical assessment shows profound linked selective effects from nongenic sources, with these effects being greater

than that of genic sources on the autosomes, as well as generally greater effects on the X chromosome than on the autosomes. I quantify these trends using multiple linear regression, and then I model the effects of linked selection to conserved elements across the whole of the genome. Places predicted to be neutral by my model do not, unlike the vast majority of the genome, show these linked selective effects. This demonstrates that linkage to these regulatory elements, and not some other mechanistic force, accounts for our findings. Further, neutrally evolving regions are extremely rare (~1%) in the genome, and despite generally larger linked selective effects on the X chromosome, the size of this "neutralome" is proportionally larger on the X chromosome than on the autosomes.

To account for this and to extend my findings to other great apes I improve on my procedure to find neutralomes, and apply this procedure to the genome of humans, Nigerian chimpanzees, bonobos, and western lowland gorillas (**Appendix B**). In doing so I show that like humans, these other apes are also enormously impacted by linked selection, with their neutralomes being substantially smaller than the neutralomes of humans. I then use my genomic predictions on neutrality to see how the landscape of linked selection changes across the X chromosome and the autosomes in regions close to, and far from, genes. While I had previously demonstrated the linked selective forces near genes are stronger on the X chromosome than on the autosomes in these taxa, I show that regions far from genes show the opposite; regions far from genes show more selection from noncoding targets on the autosomes than on the X chromosome. This finding is replicated across our great ape samples. Further, inferences on the relative effective population size of the X chromosome and the autosomes both near and far from genes can be biased as a result.

I. INTRODUCTION

Problem statements and literature reviews

To gain a deeper understanding of the contrasting roles that natural selection, genetic drift, rates of mutation and recombination play in shaping the genomic landscape one needs to gain a deep understanding of theory; both of population genetics and of computation. My dissertation reflects this principal. It begins with a formal problem in computer science; the *k*-nearest neighbors (*k*-NN) problem in generalized metric spaces. The *k*-NN problem has broad and direct applications in genomics. In genomics, proximity search can be framed as read mapping and genome alignment problems, and like many tools in machine learning, it can be used in classification. While solutions to the *k*-NN problem may in theory be extended to problems in genomics, algorithms that solve for the *k*-NN are often prohibitively slow. This chapter of my dissertation introduces the *dispersion tree*, a data structure for *k*-NN search that I hope will help lift this computational barrier.

The k-nearest neighbors problem

The *k*-NN problem is a generalization of Knuth's famous post-office problem (Knuth, 1973). In Knuth's presentation a residence can query the location of its closest post office, or 1-nearest neighbor, in the 2-dimensional space of post offices. The *k*-NN problem generalizes Knuth's presentation to *k*-closest neighbors, post offices or otherwise. More formally, the *k*-NN problem is:

- Input: A finite dataset *D* of points, a distance function *d* between points, a query *q*, and a positive integer *k*.

- Output: the k points $p_1, p_2, \ldots p_k \in D$ whose distance to q is smallest.

Solutions to the *k*-NN problem are seen in a variety of applications. When the space being considered is Euclidean, then the *k*-NN can literally be used as Knuth had originally intended, such as with nearest-neighbor search that is natively implemented in PostgreSQL. The *k*-NN can also be used in more generalized feature spaces, where close neighbors can be used as a solution for classification (Cover and Hart, 1967) and regression (Altman, 1992) problems.

The *k*-NN of a given query *q* over some dataset of points *D* can be solved in expected *O*(|D|) time and space using classic order statistic algorithms such as *quickselect* (Hoare, 1961). As both the number of queries and the number of points in such datasets can grow to be quite large, there is great need for methods that instead operate in sublinear time. While this is not possible for an arbitrary distance function *d* between the points *p* in *D,* placing restrictions on *d* may permit an organization of *D* into a data structure that can then be queried quickly. When *d* is restricted to Euclidean spaces, data structures such as *R*-trees (Guttman, 1984) are effective search tools, especially in lower dimension. The ordered pair <*d*,*D*> can also be constrained to a metric space, which is more generalized than Euclidean space. Specifically, <*d*,*D*> is a metric space if and only if, $\forall\, x, y, z \in D$:

1. $d(x, y) \geq 0$ (*non-negativity*)
2. $d(x, y) = 0$ if and only if $x = y$ (*identity of indiscernibles*)
3. $d(x, y) = d(y, x)$ (*symmetry*)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (*triangle inequality*).

Quasimetric spaces, on the other hand, satisfy these axioms excepting the rule of symmetry. To simplify measures of time, data structures in metric spaces often measure performance in units of distance computations, that is, the number of times that *d* was invoked. While this may

neglect important running-time issues such as cache-coherence, it is a convenient and machine-independent measure of time.

There are a great many spaces used in a variety of disciplines that are metric spaces. With *n*-dimensional vectors, Euclidean, Manhattan spaces are metric spaces, over strings the unit edit distance and Hamming distance are me-trics. Even $\sqrt{1 - |r(x, y)|}$, where *r* is the Pearson correlation coefficient between vectors *x* and *y*, is a metric. More problematic for issues of theory is the uniform or trivial metric, in which $d(x, y) = 1 \; \forall x \neq y$ and $d(x, x) = 0 \; \forall \; x$. The trivial metric meets the constraints of a metric space, however, without the benefit of knowing that the space is indeed a trivial metric space, searching such a metric space will in the worst-case result in an exhaustive search of the dataset to find the *k*-NN of q.

Metric data structures organize points in a metric spaces with the objective of reducing search times. Pruning is the most common technique used to reduce search-times, with all types of pruning using the distance to some number of reference points or centers to get a lower bound on the distance of a set of points to a given query. When that lower-bound is sufficiently far away, then the entire set can be discarded. There are three types of pruning: ball pruning, shell pruning and hyperplane pruning.

Ball pruning applies the concept of a "ball" to metric spaces. Balls are defined over a set of points *S* with respect to some center points *c* (*c* may or may not be a member of *S*) and have a nonnegative radius R_{max}. R_{max} is defined as $\max_{s \in S} d(s, c)$, and we define s_{max} to be an argmax of this function. Ball pruning uses a lower-bound on the distance from *q* to any point in *S*, where if that lower-bound is sufficiently large (say farther away than your current *k*-NN distance), then the ball can be pruned. Such lower bounds can be obtained by:

- $d(q,c) \leq d(q,s_{max}) + d(s_{max},c)$ $(triangle\ inequality)$
- $d(q,c) - d(s_{max},c) \leq d(q,s_{max})$
- $d(q,c) - R_{max} \leq d(q,s_{max})$

Thus, s_{max} cannot be within distance r of q if $d(q,c) - R_{max} > r$. Further, if r is a kth nearest neighbor distance, s_{max} cannot be a strictly closer neighbor of q if $d(c,q) - R_{max} \geq r$. Also note that if this property holds for s_{max}, then as it is maximal, it must hold for all points in S.

While ball pruning applies when the ball is sufficiently far from the query, shell pruning occurs when the ball is sufficiently close. We take the same definition of ball as before, save for $c \notin S$. We then take R_{min} to be the $\min_{s \in S} d(c,s)$ and s_{min} to be an argmin of this function.

- $d(c,s_{\min}) \leq d(c,q) + d(q,s_{min})$ $(triangle\ inequality)$
- $d(c,s_{\min}) - d(c,q) \leq d(q,s_{min})$
- $R_{min} - d(c,q) \leq d(q,s_{min})$

As before, s_{min} cannot be within distance r of q if $R_{min} - d(q,c) > r$, and if r is a k-NN distance, s_{min} cannot be a k-NN of q if $R_{min} - d(q,c) \geq r$. And as before, as R_{min} is minimal, if this inequality holds for s_{min} is must also hold for $\forall_{s \in S}$. Of note, both shell and ball pruning solely require the triangle inequality, making them suitable for quasimetric spaces as well as metric spaces. However, to apply both types of pruning all points must be compared to the center twice, as shell pruning uses $d(q,c)$ and ball pruning uses $d(c,q)$. Also note that a ball or a shell may have multiple centers, with the existence of lower-bound pruning on any one of these centers being sufficient to prune S.

Hyperplane pruning use not one but two centers, c_1 and c_2. With hyperplane pruning, points are partitioned into a "hyperplane," where the blocks of a partition B_1 and B_2 correspond

to the centers c_1 and c_2, with point p being assigned to the block corresponding to the argmin of $\min_{c \in c_1, c_2} d(c, p)$ and ties being broken arbitrarily. With *k*-NN search we can consider the current set of *k*-NN (i.e., the *k*-closest points encountered so far). Wlog, if $q \in B_1$, we ask the question can the *k*-NN of $q \in B_2$? Specifically, if:

1. $UpperBound[d(c_1, knn\ of\ q)] < LowerBound[d(c_2, knn\ of\ q)] \rightarrow$

   $d(c_1, knn\ of\ q) < d(c_2, knn\ of\ q) \rightarrow knn\ of\ q \in B_1 \rightarrow knn\ of\ q \notin B_2$

With the upper bound following from:

2. $d(c_1, knn\ of\ q) \leq d(c_1, q) + d(q, knn\ of\ q)$ $(Triangle\ Inequality)$

And the lower bound following from:

3. $d(c_2, q) \leq d(c_2, knn\ of\ q) + d(knn\ of\ q, q)$ $(Triangle\ Inequality)$
4. $d(c_2, q) - d(knn\ of\ q, q) \leq d(c_2, knn\ of\ q)$

Thus, after replacement in 1., pruning of B_2 occurs if:

- $d(c_1, q) + d(q, knn\ of\ q) < d(c_2, q) - d(knn\ of\ q, q)$

Note that hyperplane pruning requires both the distance of the *k*th nearest neighbor to the query, and the distance of the query to its *k*th nearest neighbor. As such, hyperplane pruning is usually restricted to symmetric distance computations as that halves the number of such computations. Also note that hyperplane partitioning may cause a constant number of points to be assigned to one of the blocks, causing imbalance in the resulting data structure. Adding a factor ε can be used to balance the partition. For example, partition assignments can be in accordance to $\min[d(c_1, p), d(c_2, p) + \varepsilon]$, though the pruning rules would need to incorporate ε to ensure correctness. Also note that hyperplanes can easily extended to *m* centers, where the points are

partitioned in accordance to whichever center they are closest to, and pruning occurring with respect to every pair of centers.

There are two basic subtypes of exact algorithms that solve the *k*-NN problem: Matrix based-approaches, and tree-based approaches. Matrix-based approaches are largely a reformulation of the approximation and elimination search algorithm (AESA) (Vidal 1986) and further refinements by Micó et al. (1994). While first introduced as a solution to the 1nn problem, these approaches can be extended to *k*-NN (Aibar et al. 1993; Morena-Seco et al., 2002). With the AESA, every point p in D is compared to every other point, and the exact distance between them is stored, with this information being stored in a matrix M of size $\Omega(n^2)$, where $n = |D|$. Then, a single distance computation is performed to an arbitrary point p_1 which becomes the first candidate *k*-NN. p_1 can then be seen as a center, and every other point, when coupled with this center, can be seen as either a ball or a shell. An array of maximal lower bounds B of size n is then initialized to the maximum lower bound from ball or shell pruning with respect to p_1 using the distances in M. Thus the bounds in B serve to approximate their distance to the query, and the points that are sufficiently far away (that is, farther away than the *k*-NN distance) can be eliminated. The next point chosen, p_i, is the argmin of B; if B[i] is farther away than our current *k*-NN, it follows that p_i can be pruned, and as p_i is minimal it follows that the rest of D can be pruned as well, and the search terminates. Otherwise p_i cannot be pruned, so a distance computation is spent comparing it to q, and then all points that cannot be pruned have their bounds in B updated to contain maximum lower bounds using M and pruned as appropriate. With linear-AESA (Micó et al. 1994), a constant number (m) of base prototypes b are chosen, where instead a precomputed matrix of size $\Omega(nm)$ is created. With Micó et al. (1994), B is

initialized using an arbitrary point in b, and updates to B do not use the entire dataset, but just the m reference points and only in the case that the chosen point, p_i, was a reference point.

Several properties of the AESA make it unsuitable for many instances of the nearest-neighbor problem. First observe that while the AESA and LAESA are argued to invoke a constant number of distance computations (taking k as a constant), even in the case that p_1 is the true *1*nn in a *1*nn search, the entirety of B is searched. Though this search requires only a single distance computation, this exhaustive search implies that AESA and LAESA operate in $\Omega(n)$ time. Second, AESA in particular requires $\Omega(n^2)$ memory and LAESA requires $\Omega(mn)$ memory. While the quadratic memory requirements of AESA are a substantial barrier to its broad applicability, even the treatment of m as a constant in LAESA may be overly-optimistic. As n goes to infinity, it may be that m may need to be a sublinear function of n to continue to generate k-NN with a constant number of distance computations. Both AESA and LAESA are applicable in some domains, for example when the distance function is extremely costly and/or when the size of the datasets is small. They further exemplify how the reliance on distance computations as a measure of time can be exploited by k-NN search methods to exaggerate their apparent speeds. AESA and LAESA are also the first algorithms to use lower-bound k-NN search, a search strategy that does no more distance computations than a range search started with the true k-NN distance (Hjaltason and Samet, 2000).

The vast majority of metric space indexes are tree-based. First introduced for discrete metrics (Burkhard and Keller 1973), metric trees were more generally introduced by Uhlmann (1991). Uhlmann (1991) presented both ball trees and generalized hyperplane trees. Ball trees

recursively partition points in the metric space into two blocks, a ball block and a shell block, with respect to an arbitrary center *c* as follows:

- Input: A dataset *D* of points, and a distance function *d*
- Output: A ball tree formed over *D* with respect to *d*

Procedure BallTree(D, d)

1. If *D* is empty, return null
2. c := an arbitrary point in *D*
3. *m* := median distance over all points in *D* with respect to *c*
4. node.center := c
5. node.median := m
6. node.ball := BallTree($\forall_{D_i \in D} | \, d(c, D_i) \leq m$, d)
7. node.shell := BallTree($\forall_{D_i \in D} | \, d(c, D_i) \geq m$, d)
8. return node

Ties at the median value *m* are broken as to ensure that the ball and the shell are balanced. Ball trees are thus balanced search trees, which use ball pruning and shell pruning as appropriate. Yianalos (1993) introduced the vantage-point tree, which further refined ball trees by using heuristics to select centers that maximize the variance in *d*. Multi-vantage point (MVP) trees represent another such refinement, where multiple centers are chosen, as well as multiple order statistics (instead of just medians), and minimum and maximum radii of each of the blocks with respect to each of the centers, allowing both ball-pruning and shell-pruning of any of the blocks (Bozkaya and Ozsoyoglu 1999).

Uhlmann (1991) also introduced generalized hyperplane (GH) trees. GH trees form a hyperplane partition between two centers, with each block being recursed upon to create a tree. While Uhlmann (1991) did not specify the center-finding criteria, hyperplane pruning will be more likely if the centers are far apart. Brin (1995) extends GH trees to geometric near neighbor access trees (GNAT). GNATs have an arity *a*, and form a hyperplane partition using *a* centers. GNATs use a max-min approach for selecting centers; they start with an arbitrary point, then select the point in the dataset farthest from that, and they then select subsequent points whose minimum distance to any of the centers chosen so far is maximal. Spatial approximation trees (Navarro 2002) are a type of GNAT where the centers are chosen to be closer to some center than to any other point, while antipole trees (Cantone et al. 2005) are a mixture of hyperplane trees (near the root) and ball trees (near the leaves).

Cover trees (Beygelzimer et al. 2006) are another ball-tree that provide an interesting theoretical perspective on *k*-NN search in metric spaces. While the previous ball trees reduce the number of points at each level of recursion by some constant factor (2 in the case of vantage point trees), cover trees are level trees that reduce the radius of balls at each level by a constant factor (1.3 in practice, though 2 is used in their proofs) at each level. For level *i* of a cover tree, the minimum distance between any pair of points is $> 2^i$, with the root level occurring at level $+\infty$ and the leaves at level $-\infty$. Cover trees assume that the metric space is constrained to have an expansion constant *c*, where *c* is minimal over $\forall_r \forall_{p \in D}\ |B(p, 2r)| \leq c|B(p, r)|$, where $|B(p, r)|$ is the number of points in ball *B* centered on point *p* with radius *r*. Note that if the points are drawn from a uniform Euclidean space of dimension *d*, then $c \sim 2^d$. With this design, cover trees can be constructed in time $O(c^6 n \log n)$ and they can be queried in time $O(c^{12} \log n)$. Cover trees are

one of the few metric data structures that make asymptotic claims on *search* in metric spaces, though the constants (which as noted by Beygelzimer et al. 2006 may not be constant in practice) are large enough that they may dominate search times even when the size of the dataset is large.

While cover trees provide a rich theoretical approach to search in metric spaces, the list of clusters (Chávez and Navarro 2000) is an exceedingly simple data structure that has surprisingly fast search-times in practice. With the list of clusters, *m* centers are chosen, and then balls are greedily formed each with $\frac{n}{m}$ points. The balls are then placed in a list, creating a list of balls. Further refinements of the list of clusters just use a hyperplane partition using *m* centers (Tellez and Chávez 2012), greatly improving construction times, though the blocks of the partition may not each have $\frac{n}{m}$ points. The points within each ball are *not* organized, and are searched exhaustively if they cannot be pruned. While *m* can be any value, the authors suggest setting *m* to $\sqrt{2n}$. Despite this simplicity, the list of clusters can be a very competitive search strategy in higher dimension. Of note, the list of clusters is the only data structure that does not create partitions with respect to a constant number of centers.

Diversity in the genome.

We turn now to problems in population genetics and genomics. My dissertation uses high coverage whole genome data from a sample of two human populations (**Appendix A**), as well as samplings in humans and other great apes (**Appendix B**) to evaluate the effects of "selection at linked sites" (i.e., linked selection) on genomic patterns of diversity.

A common estimator of genetic diversity at the nucleotide level is the index of nucleotide diversity (π) (Nei and Li 1979). π is the average number of pairwise differences between

individuals in a population, and under neutral equilibrium conditions π is an estimator of the population mutation rate Θ, where $\Theta_\pi = 4Ne\mu$. Thus π is influenced both by the (individual) mutation rate (μ), and the effective population size (Ne), as well as by violations of neutral equilibrium conditions.

Many processes mechanistic processes influence π through changing μ. Examples include DNA replication timing (Koren et al. 2012, 2014), trinucleotide context (Hwang and Green 2004), paternal age (Kong et al. 2012; Venn et al. 2014), and recombination itself (Pratto et al. 2014; Arbeithuber et al. 2015). Ne, too, changes, both genome-wide and locally. Genome-wide changes are due to demographic effects such as migration, population structure, and population size changes. At local scales, the indirect effects of selection on linked neutral sites also reduces Ne (Maynard Smith and Haigh 1974; Charlesworth et al. 1993; Hudson and Kaplan 1995). Linkage to positively selected alleles also skews the allele frequency spectrum (AFS) (Braverman et al. 1995), and when the magnitude of selection is small, negative selection behaves similarly but to a lesser degree (Williamson and Orive, 2002; Nicolaisen and Desai 2012, 2013).

To disentangle the effects of mutation rate from that of effective population size π can be divided by divergence (D). While π is the average number of differences between ingroups, D is the (typically average) number of differences between ingroups and an outgroup. The expected value of D is $2\mu t + 4Na\mu$, where t is the speciation time between the ingroups and outgroups and Na is the effective population size of the ancestral taxon. Thus, the expected value of D is equal to μ times a constant. Taking the ratio π/D transforms an estimator of Θ to an estimator of Ne that controls for the local mutation rate.

While π/D serves as an estimator of Ne, this estimator may still be confounded by mutation rate in some cases. Taking recombination as an example, recombination is itself mutagenic (Arbeithuber et al. 2015). In humans, 80% of recombination events occur in <10% of the genome (McVean et al. 2004) forming so-called recombination "hotspots," a finding that is generally recapitulated across great apes (Stevison et al. 2016). Hotspots are not a conserved feature of the genomic landscape; they are neither conserved between humans and chimpanzees (Ptak et al. 2005), nor or they necessarily conserved between people of West African ancestry and European ancestry (Hinch et al. 2011). Further, recombination is associated with gene conversion; while recombination resolves double-strand breaks with a cross-over event, the more common resolution of the breakpoint is with a gene-conversion event. Gene conversion and recombination both involve a strand-invasion phase, where the maternal and paternal chromosomes are paired. Mismatches in this pairing are resolved with a bias against weakly bonded base-pairs (W: A,T), favoring instead strongly bonded pairs (S: G,C), with an effect known as GC-biased gene conversion (gBGC) (Marias, 2003). This bias acts as a weak selective pressure (Galtier et al. 2009), and as gene conversion and recombination are spatially correlated, it is another example of how recombination and diversity may be associated.

Like recombination, DNA replication timing is another mechanistic process that may also disproportionally affect π more than D. DNA replication occurs in parallel, with some regions replicating earlier than others. Early-replicating regions have lower rates of mutation, both of soma and of germlines, than late-replicating regions (Koren et al. 2012). The timing of DNA replication is tightly regulated, with this regulation being modulated in *cis* by different alleles in human populations (Koren et al. 2014). Thus it follows that local mutation rates likely also vary

between populations because of this difference, which suggests that like recombination, differences in replication timing may affect diversity more than divergence. Thus, while dividing π by D controls for some aspects of mutation-rate heterogeneity, if the properties that influence mutation-rate are transient, this control may be incomplete.

Natural selection may also have profound effects on π/D. These effects can be partitioned into two sets; effects that are direct, and based on the action of selection on genomic substrates, and effects that are indirect, and stem from the action of a selected allele on the genealogy of linked sites. Most studies in humans argue for a large role of negative selection, with lesser degrees of positive directional selection (Akey 2009; Lohmueller et al. 2011; Hernandez et al. 2011; Enard et al. 2014), and even lesser degrees of other types of selection such as balancing selection (Leffler et al. 2013).

Negative selection purges deleterious mutations from populations. When purged, both the site of selection and nearby sites are removed in an effect known as background selection. Charlesworth et al. (1993) first introduced a model that describes the action of negative selection on the genealogy of linked sites. When deleterious alleles are purged quickly from the population the effect at linked sites is equivalent to a reduction in the local Ne (Hudson and Kaplan 1995); that is, mean coalescent times are reduced but the structure of the genealogy remains the same. As shown by both theoretical and forward-in time simulations, linkage to weakly deleterious alleles, on the other hand, reduces Ne and skews the underlying genealogy (Fu 1997; Williamson and Orive, 2002; Nicolaisen and Desai, 2012, 2013). Specifically, individuals with many deleterious alleles in the past are less likely to contribute alleles in the present, which distorts how segregating sites are distributed on the genealogy. And since purifying selection has been

more effective in removing alleles in the distance past than in the present, background selection emulates population growth by increasing the terminal branch lengths. Background selection reduces diversity both as a function of the local deleterious mutation rate, *u*, and the rate of recombination *r*. In the model of Husdon and Kaplan (1995), the reduction in Ne from background selection is $e^{-\frac{u}{2sh+r}}$ for neutral loci *r* genetic units away from the selected site, though the selection coefficient (s) and dominance coefficient are sometimes neglected as they can be much smaller than *r*.

Positive directional selection drives alleles to fixation. With the hitchhiking effect, when a selected allele goes to fixation alleles on the same genetic background "hitchhike" to fixation (Maynard Smith and Haigh 1974). In 2-locus hitchhiking models, heterozygosity at a single neutral allele at the time of fixation largely depends on the ratio *s*/*r*, where *s* is the selection coefficient and *r* is the genetic distance between the neutral and selected allele (assuming 2Ne*s* is large, e.g. > 100) (Maynard Smith and Haigh 1974; Kaplan et al. 1989). The model of Maynard Smith and Haigh (1974) has been extended from single to recurrent hitchhiking events, where hitchhiking spontaneously may occur throughout the chromosome (Kaplan et al. 1989). While background selection primarily reduces Ne and has a relatively small effect on low-frequency (predominantly singleton) polymorphisms (Fu 1997), hitchhiking skews both low- and high-frequency polymorphisms (Braverman et al. 1995; Fay and Wu 2000), with hitchhiking's effect on high-frequency variants stemming from recombinant haplotypes containing the selected allele also being driven to fixation.

To understand the roles that genetic hitchhiking and background selection have on genomic patterns of diversity we need to have a firm understanding how both these processes

behave, and of where the targets of selection are in the genome. Genes are one obvious class of elements that may be under selection, and their distribution in the genome is largely known (but see Nelson et al. 2016). Phylogenetically conserved elements are another likely target. Phylogenetically conserved elements are genomic locations that, when constrained by a phylogeny, have fewer substitutions than would otherwise be predicted. When these elements are identical between species such as humans and mouse, they are termed "ultraconserved" (Bejerano et al. 2004). Ultraconserved elements are not just mutational "coldspots," but instead show extreme skews in the AFS towards an excess of rare variants, consistent with negative selection (Katman et al. 2007). *phastCons* conserved elements are another class of constrained elements (Siepel et al. 2005). With *phastCons* elements, conserved elements are predicted by a phylogenetic hidden Markov model (phyloHMM). The *phastCons* phyloHMM has a constrained and an unconstrained model state, and transition probabilities between (and within) these states. Both states depend on multiple sequence alignment given a known phylogeny, where the time-scales of the constrained phylogeny are reduced by a constant factor from the unconstrained phylogeny. *phastCons* elements are then elements that are better explained by the constrained versus the unconstrained phylogeny. *phastCons* elements are another class of constrained elements in the genome, and like ultraconserved elements, these elements appear to be under both positive and negative selection (Halligan et al. 2011, 2013).

A related concept to conserved elements are functional elements. While selective elements are computationally inferred, functional elements are instead ascertained using biochemical methods. Thus, it follows that while elements such as *phastCons* elements are likely targets of natural selection, biochemically inferred elements such as loci identified by ChIP-Seq

and GRO-Seq may (Yu et al. 2015), or may not be under appreciable levels of selection (Kellis et al. 2014). Further, biochemically determined functional elements require explicit knowledge of which transcription factors to assay in which cells and under which conditions. These functional elements thus may represent a superset of selective elements, as they all may not be selective, and a subset of elements, as the proper cells in the proper conditions may not have been surveyed (e.g. Ostuni et al. 2013). Conserved elements, however, are inferred in a manner agnostic to function and indeed these elements serve a variety of biological roles. Some conserved elements are aberrantly expression in cancer cells (Braconi et al. 2011). In genes, ultra-conserved regions affect mRNA secondary structure (Sathirapongsasuti et al. 2011), regulate alternative splicing (Lareau et al. 2007; Ni et al. 2007; Sathirapongsasuti et al. 2011). Conserved regions also serve as tissue-specific enhancers (Pennacchio et al. 2006). As such, conserved regions are a heterogenous mixture of genomic elements with the sole commonality being conservation.

As I have outlined above, recombination and diversity are intimately associated. The correlations between them may be owed to cellular processes such as GC-biased gene conversion and recombination's mutagenicity, or they may be owed to unlinking selected alleles from the genomic background. Further, in the case of selection it may be that linked selection may stem from linkage to genic sites of selection, or it may stem from linkage to nongenic elements. Thus, in the studies that follow I seek to disentangle the mechanistic from the selective, and to highlight whether or not selection to genes or to non-coding elements drive genomic patterns of diversity in both humans and other great apes.

II. PRESENT STUDY

The appendices to this thesis describe in detail the background, methods, results, discussion that make up my dissertation. I give in this section the major findings and novel advancements of these appendices, beginning with the finding that most bases in the human genome are impacted by natural selection in the paper for **Appendix A** entitled:

THE ROLE OF PHYLOGENETICALLY CONSERVED ELEMENTS IN SHAPING THE LANDSCAPE OF HUMAN GENOMIC DIVERSITY

examines the extent of linked selection in the human genome. I present an empirical assessment of the effects of selection at linked sites in the human genome. I show that these effects are pervasive, occurring both near and far from genes, leaving few nucleotides unaffected by these linked effects. Using masking and two datasets of *de novo* mutations I show that these effects are not due to GC-biased gene conversion, nor are they due to recombination's mutagenicity. I use a genomic model of linked selection to *phastCons* elements, and show that regions predicted to be neutral by my model do not show these linked selective effects, which shows that linkage to *phastCons* elements drive genomic patterns of diversity. I then introduce a definition of "neutral" that has a statistical basis, and show that the regions in the genome that appear to be neutral are exceedingly rare (~1% of the genome).

In my work shown in **Appendix B**, entitled:

GENOMIC INFERENCE ON SEXUAL SELECTION IN THE GREAT APES

I extend my research from **Appendix A** to gorillas, Nigerian chimpanzees and bonobos, showing that the genomes of each of these species are as impacted by selection as humans. I then model linkage to *phastCons* elements inferred over different time scales and show that these elements

are a heterogeneous class, and further, they have substantially different distributions on the X chromosome and the autosomes. I then re-applied my model of linked selection to *phastCons* elements in great apes, and introduce a general framework for generating "neutral" regions. I also show that while regions of the X chromosome near genes show more (linked) selective effects than the autosomes, regions far from genes show the opposite. I compare estimates of Ne on the X chromosome versus the autosomes, and show that the relative effective population sizes far from genes are biased by this variable level of on nongenic elements. I finish by estimating the breeding sex ratio in African great apes, and show that this ratio is especially elevated in chimpanzees. I posit that either sperm competition arose in the lineage of *Pan troglodytes*, and not in the ancestor of bonobos as chimpanzees, or that behavioral strategies in chimpanzees especially elevate the breeding sex ratio in chimpanzees compared to bonobos.

In my work shown in **Appendix C**, entitled:

FASTER METRIC NEAREST NEIGHBOR SEARCH USING DISPERSION TREES.

I describe a data structure called a dispersion tree which is an index of a metric space. Dispersion trees can be used to solve both *k* nearest-neighbor searches, as well as range searches. Dispersion trees take some of the principals that I gleaned from other metric indexes, and extend them in important ways. Nearly all metric indexes are recursively constructed top-down, with the points in the space being partitioned in accordance to their distance to some constant number of references or centers. High dimensional spaces, metric or otherwise, have the unfortunate property that few points are close, and most points are generally far apart. Thus, it follows that partitioning strategies that use a constant number of references will place points that are far from all references into blocks that are largely arbitrary, an effect that is likely exacerbated with each

recursive call. Conversely, data structures such as the list of clusters, which in practice form $O(\sqrt{n})$ centers and place points into partitions organized by these centers, have very fast search times. As the list of clusters provides little organization of the points beyond this initial partition, this speaks to how the top-down construction used by nearly every other metric space index may be affecting search. True bottom-up strategies, however, are infeasible for large datasets, as they involve the mergers over the quadratic number of pairs of points. Bottom-up procedures also need an objective function on how to evaluate the quality of these possible mergers.

With these principles in mind, I introduce dispersion trees. Dispersion trees are the first data structure in metric spaces that uses a partial bottom-up construction technique, and they have sub-quadratic construction time ($O(n^{1.5} \log n)$ to be precise) and use $O(n)$ space. Dispersion trees use $\sqrt{n}$ dispersed points, and build a search hierarchy based on this dispersed set. Dispersion trees have faster average-case search times in a variety of metric spaces than a large sampling of competing metric data structures. There are only two exceptions to this: dispersion trees are slower than comparable data structures in lower-dimensional (<6-dimensional under the Euclidean distance function when considering points in a unit hypercube) unstructured spaces, and very in high-dimensional spaces (structured or not), which generally devolve into a near-linear scan of the dataset. My work shows that bottom-up strategies are a powerful technique for constructing metric indexes, and it opens up this strategy for future research.

III. FUTURE DIRECTIONS

My work in the areas of nearest neighbor search and in the extents of linked selection open up several interesting lines of future research. With nearest neighbor search, while dispersion trees have fast query times, their construction times are slower than comparable data structures. Further, dispersion trees only use a partial bottom-up strategy. It follows that if the whole dataset could be considered simultaneously then a true bottom-up strategy could join subtrees that are even closer, resulting in smaller ball radii and thus more pruning and faster search times. Algorithms that do this quickly (e.g., $O(n\ log^x n)$ time) would present a formidable construction strategy. Further, dispersion trees use an objective function that is greedily minimal on radius; that is it merges subtrees under the objective that the merged subtree's radius is minimal over all possible pairs. However, the number of possible mergers shrinks after each merge. This may cause the initial subtrees formed to have very small radii, as there are many possible choices, but subtrees that contain many points may have (much) larger radii because they were formed over a smaller set of possible mergers and they are constrained by the choice of centers. I posit that there is an optimal *radius* for pruning, with optimal taking both the probability of pruning and the number of points pruned away into account (I currently neglect the latter property). It could also be that dispersion trees, given this reduced flexibility in larger subtrees when the subtrees are created bottom-up, may be sub-optimal with respect to this optimal radius, and that forming balls with optimal radii initially may be a better strategy.

There are several promising lines of research that could further my work on linked selection. First, better models that scale to the level of the genome are in great need. The extant genomic models of linked selection are restricted to the background selection model of

(McVicker et al. 2009) and Model C of Halligan et al. (2013). The shortcomings of the McVicker et al. (2009) model are both theoretical and practical; models of background selection that neglect weakly selected alleles may not be entirely appropriate for genomic models of linked selection. Model C also has significant shortcomings; many combinations of parameters yield very similar levels of model fit in humans and great apes. While these different best-fitting models largely agree on which loci are "neutral," assessing levels of linked selection beyond the purposes of this classification are more problematic. This issue of singularity could in part could be due to insufficient power; perhaps the loci in humans and other apes were too large to accurately infer these parameters. Or, it could be that our reliance on two classes of selective elements, CEEs and CNEs, is too coarse. As I show with my "clusters" of primate scores, what constitutes an element may be abstracted away; perhaps different bases under different levels of constraint (both positive and negative) would provide a more realistic substrate on which these models of linked selection may operate.

REFERENCES

Aibar P, Juan A, Vidal E. 1993. Extensions to the approximating and eliminating search algorithm (AESA) for finding k-nearest-neighbours. New Advances and Trends in Speech Recognition and Coding. 2328.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here?. Genome Research. 19: 711-22.

Altman NS. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46: 175-85.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proceedings of the National Academy of Sciences. 112: 2109-14.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science. 304: 1321-5.

Beygelzimer A, Kakade S, Langford J. 2006. Cover trees for nearest neighbor. Proceedings of the 23rd international conference on Machine learning. ACM. 97-104.

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci 85: 6414-8.

Bozkaya T, Ozsoyoglu M. 1999. Indexing large metric spaces for similarity search queries. ACM Transactions on Database Systems (TODS). 24: 361-404.

Braconi C, Valeri N, Kogure T, Gasparini P, Huang N, Nuovo GJ, Terracciano L, Croce CM, Patel T. 2011. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. Proc Natl Acad Sci 108:786-91.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783-96.

Brin S. 1995. Near neighbor search in large metric spaces. Proceedings of the 21th International Conference on Very Large Data Bases. 1995: 574-584.

Burkhard WA, Keller RM. 1973. Some approaches to best-match file searching. Communications of the ACM. 16: 230-6.

Cantone D, Ferro A, Pulvirenti A, Recupero DR, Shasha D. 2005. Antipole tree indexing to support range search and k-nearest neighbor search in metric spaces. IEEE Transactions on Knowledge and Data Engineering. 17: 535-50.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-303.

Chávez E, Navarro G. 2000. An effective clustering algorithm to index high dimensional metric spaces. String Processing and Information Retrieval, 2000. IEEE SPIRE 75-86.

Cover T, Hart P. 1967. Nearest neighbor pattern classification. IEEE transactions on information theory. 13: 21-7.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. Genome research. 24: 885-95.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155: 1405-13.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends in Genetics. 25: 1-5.

Guttman A. 1984. R-trees: a dynamic index structure for spatial searching. ACM. 14: 47-57.

Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. Mol Biol Evol 28: 2651-60.

Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet.9: e1003995.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. Science. 331: 920-4.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, Aldrich MC. 2011. The landscape of recombination in African Americans. Nature. 476: 170-5.

Hjaltason GR, Samet H. 2000. Incremental similarity search in multimedia databases. Technical Report 4102, University of Maryland.

Hoare CA. 1961. Algorithm 65: find. Communications of the ACM. 4: 321-2.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. Genetics. 141: 1605-17.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America. 101: 13994-4001.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 147: 915-25.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics. 123: 887-99.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. Science. 317: 915

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I. 2014. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences. 111: 6131-8.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 488: 471-5.

Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. The American Journal of Human Genetics. 91:1033-40.

Koren A, Handsaker RE, Kamitaki N, Karlić R, Ghosh S, Polak P, Eggan K, McCarroll SA. 2014. Genetic variation in human DNA replication timing. Cell. 159: 1015-26.

Knuth DE. 1973. The Art of Computer Programming, Vol. 3, Sorting and Searching. Addison-Wesley. Reading, Mass. 578-9.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 1997. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature. 446: 926-9.

Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, Donnelly P. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science. 339: 1578-82.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, Jørgensen T. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 7:e1002326.

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

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science. 304: 581-4.

McVicker GA, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:e1000471.

Micó ML, Oncina J, Vidal E. 1994. A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements. Pattern Recognition Letters. 15: 9-17.

Moreno-Seco F, Micó L, Oncina J. 2002. Extending LAESA fast nearest neighbour algorithm to find the k nearest neighbours. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 718-724. Springer Berlin Heidelberg.

Navarro, G. 2002. Searching in metric spaces by spatial approximation. The VLDB Journal. 11: 28-46.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences. 76: 5269-73.

Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, Reese AL, McAnally JR, Chen X, Kavalali ET, et al. 2016. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science. 351: 271-5.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, Ares M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev 21: 708-18

Nicolaisen LE, Desai MM. 2012. Distortions in genealogies due to purifying selection Mol Biol Evol 29: 3589-3600.

Nicolaisen LE, Desai MM. 2013. Distortions in genealogies due to purifying selection and recombination. Genetics. 195: 221-30.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499-502

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. Science. 346(6211):1256442.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nature Genetics. 37: 429-34.

Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, Curina A, Prosperini E, Ghisletti S, Natoli G. 2013. Latent enhancers activated by stimulation in differentiated cells. Cell 152: 157-71.

Sathirapongsasuti JF, Sathira N, Suzuki Y, Huttenhower C, Sugano S. 2011. Ultraconserved cDNA segments in the human transcriptome exhibit resistance to folding and implicate function in translation and alternative splicing. Nucleic Acids Res 39: 1967-79.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-50.

Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Bustamante CD, Hammer MF, Wall JD. 2016. The Time Scale of Recombination Rate Evolution in Great Apes. Molecular Biology and Evolution. 33: 928-45.

Tellez ES, Chávez E. 2012. The list of clusters revisited. Mexican Conference on Pattern Recognition. *Springer Berlin Heidelberg.*187-196.

Uhlmann JK. 1991. Satisfying general proximity/similarity queries with metric trees. Information processing letters. 40: 175-9.

Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, McVean G. 2014. Strong male bias drives germline mutation in chimpanzees. Science. 344: 1272-5.

Vidal, E. 1986. An algorithm for finding nearest neighbours in (approximately) constant average time. Pattern Recognition Letters. 4: 145-57.

Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. Molecular biology and evolution. 19: 1376-84.

Yianilos PN. 1993. Data structures and algorithms for nearest neighbor search in general metric spaces. SODA. 93: 311-21.

Yu F, Lu J, Liu X, Gazave E, Chang D, Raj S, Hunter-Zinck H, Blekhman R, Arbiza L, Van Hout C, et al. 2015. Population Genomic Analysis of 962 Whole Genome Sequences of Humans Reveals Natural Selection in Non-Coding Regions. PLoS One 10: e0121644.

APPENDIX A

THE ROLE OF PHYLOGENETICALLY CONSERVED ELEMENTS IN SHAPING THE LANDSCAPE OF HUMAN GENOMIC DIVERSITY

Manuscript in preparation for submission to Molecular Biology and Evolution

Intended as a research article at Molecular Biology and Evolution

TITLE

The role of phylogenetically conserved elements in shaping the landscape of human genomic diversity

Keywords:
Phylogenetic conserved elements
Recombination
Linkage
Selection,
Diversity
Null model

August E. Woerner[1], Krishna R. Veeramah[2], Joseph C. Watkins[3], and Michael F. Hammer[1*]

[1]*ARL Division of Biotechnology, University of Arizona, Tucson, AZ USA, 85721*

[2]*Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY USA 11794*

[3]*Department of Mathematics, University of Arizona, Tucson, AZ USA, 85721*

**Correspondence should be addressed to:*

Michael F. Hammer, PhD

ARL Division of Biotechnology, 111K Keating Building, 1657 E. Helen Street, University of Arizona, Tucson, AZ, 85721, USA

Phone: +1-520- 621-9828

e-mail: mfh@email.arizona.edu

ABSTRACT

Evolutionary genetic studies in many species have shown a positive correlation between levels of nucleotide diversity and rates of recombination. More recently, diversity has been shown to be positively correlated with the genetic distance to genes in humans and great apes. Both positive-directional and purifying selection have been offered as the source of these correlations via genetic hitchhiking and background selection, respectively (here referred to as linked selection). Phylogenetically conserved elements (CEs) are short (~100bp), widely distributed sequences (comprising ~5% of genome) that, are often found far from genes. While the function of many CEs is not known, studies in hominids and murids have provided evidence that CEs also are associated with reduced diversity at linked sites. Here we contrast diversity both with distance to genes and with rates of recombination, parsing the effects of linked selection into genic and nongenic sources. Using high coverage (>80x) whole genome data from two human populations, the Yoruba and the CEU, we perform fine scale evaluations of diversity, rates of recombination, and linkage to genes. We find that the local rate of recombination has a stronger effect on levels of diversity than linkage to genes on the autosomes, and that these effects of recombination persist even in regions far from genes. Notably, low recombination-rate loci that are far from genes have ~16% less diversity than high recombination-rate loci close to genes. The mutagenic effects of recombination and GC-biased gene conversion are unlikely to contribute to the correlation between diversity and the local recombination rate. On the other hand, our whole-genome modeling shows that selection on sites within or linked to CEs is consistent with this finding. A major implication of this result is that very few sites in the genome are predicted to be free of the effects of selection. These sites, which we refer to as the human "neutralome", comprise only 1.2% of the autosomes and 5.1% of the X chromosome. Demographic analysis of the neutralome reveals larger population sizes and lower rates of growth for ancestral human populations than inferred by previous analyses.

INTRODUCTION

Recombination is a fundamental force of molecular evolution. Nearby sites on the chromosome are likely co-inherited, implying a correlation in their evolutionary trajectory. This linkage disequilibrium (LD) between sites reduces the efficiency of natural selection, decreasing fixation rates for positively selected alleles while increasing said rates for negatively selected alleles (Hill and Robertson 1966). Selected alleles may be in LD with neutral alleles. Neutral alleles can hitchhike to fixation under a selective sweep (Maynard Smith and Haigh 1974) or to extinction under background selection (Charlesworth et al. 1993). Both background selection and hitchhiking distort the underlying ancestral recombination graph (ARG), increasing terminal branch lengths, altering the allele frequency spectrum (AFS) and generally reducing genetic diversity (Braverman et al. 1995; Tachida 2000; Williamson and Orive 2002; O'Fallon et al. 2010; Nicolaisen and Desai 2012, 2013). We refer to these processes as linked selection.

Recombination decouples selected alleles from the genomic background, mitigating the effects of linked selection. In their landmark study, Begun and Aquadro (1992) found that the rate of recombination (R) is significantly correlated with nucleotide diversity (π), a finding that has been replicated and expanded upon in a variety of contexts (Nachman 2001; Spencer et al. 2006; Cai et al. 2009; Lohmueller et al. 2011; McGaugh et al. 2012). The implications of this body of work are that linked selection plays a major role in constraining diversity across the genome. Similarly, diversity and the minimum genetic distance to genes (G) is also positively correlated (Hammer et al. 2010; Gottipati et al. 2011; Prado-Martinez et al. 2013; Arbiza et al. 2014). Cellular processes such as GC-biased gene conversion (Marias

2003) (but see McGaugh et al. 2012) or the mutagenicity of recombination (Pratto et al. 2014; Arbeithuber et al. 2015; Francioli et al. 2015) may also contribute to correlations between R and π.

An important question that has not been carefully addressed in humans is how much of the effects of linked selection is driven by evolutionarily constrained regions within and outside of genes. Phylogenetically conserved elements are loci that have far fewer substitutions than would otherwise be expected in a neutral region over a given phylogeny. Classes of phylogenetically conserved elements include ultraconserved elements, which within humans are defined as having 100% identical orthologs between humans, mouse and rat (Bejarno et al. 2004). More relaxed definitions of conserved elements, such as *phastCons* elements (Siepel et al. 2005), permit low levels of nucleotide substitution. Conserved elements are not simply mutational coldspots, and instead appear selective (Bejarno et al. 2004; Cooper et al. 2005; Siepel et al. 2005; Katzman et al. 2007; Chen et al. 2007; McVicker et al. 2009; Halligan et al. 2011). In particular, diversity is reduced within and adjacent to both coding and noncoding *phastCons* elements (Hernandez et al. 2011; Halligan et al. 2013). Unlike genes, conserved elements are generally short (~100bp) and widely distributed sequences (comprising ~5% of the genome), suggesting that linked selection to phylogenetically conserved elements may also play role in constraining diversity across the genome.

We investigate the relationships between π, R, and G at genic and nongenic sites across the genome, making use of high coverage whole genome sequence data from two human populations: the Yoruba from Ibadan, Nigeria, and Northern Europeans from Utah. We also test the hypothesis that recombination is mutagenic within the non-repetitive portions of the genome that we consider in our estimates of diversity, and assess the extent to which GC-biased gene conversion contributes to the relationships among diversity, R and G. We model linkage to phylogenetically conserved sequences to assess their role in shaping genomic patterns of diversity, and then validate this model by testing for

linked selection in regions predicted to be unlinked to *phastCons* elements. This leads us to a hypothesis-testing framework and a test of "neutrality" for genomic regions.

RESULTS

Diversity and linkage on the X chromosome and the autosomes

In order to estimate local genetic diversity we partitioned the human genome into non-overlapping 10kb loci. After masking out several classes of repeats, we computed nucleotide diversity (π) and divergence (D) to a human-orang ancestor sequence within each locus for both the Yoruba (YRI, n=9) and Northern Europeans from Utah (CEU, n=9) based on high coverage (~80x) whole genomes. Next, we assessed the effects of linkage on diversity at each locus by computing both the probability that the locus is decoupled from the closest gene (G), i.e., the minimum genetic distance to genes in centimorgans (cM), and the probability of recombination within the locus (R), i.e. the length of the locus in cM (Supplemental Figure S1).

To visualize how G and R jointly affect the local effective population size (i.e. π/D), we bin loci using the marginal distributions of both G and R, by deciles in the autosomes and quartiles on the X chromosome. For each bin, we display the median value of π/D (division by D controls for variation in local mutation rate amongst loci) (**Figure 1**, Supplemental Figure S2). As expected, we find that diversity increases with the genetic distance to genes (left to right columns), consistent with previous work (Hammer et al. 2010; Gottipati et al. 2011; Arbiza et al. 2014). However, surprisingly π/D also increases with R (bottom to top rows) for both types of chromosomes regardless of the value of G. This gradient is even evident in the last column of **Figure 1**, which consists of loci that are essentially unlinked from all genes. Both π (Supplemental Figure S3) and D (Supplemental Figure S4) examined separately show the same general trends, though with a much more pronounced effect for π.

In order to assess statistical significance of the observed trends we used iterated re-weighted least squares (IRLS) regression on the model $\pi/D \sim G + R$ for autosomes and the X chromosome separately, with estimated slope coefficients β providing a measure of effect size. The standardized slope coefficients β are significantly greater than 0 ($p<0.001$ across all comparison). Further they satisfy $\beta_{XG} > \beta_{XR} > \beta_{AR} > \beta_{AG}$ ($p < 0.001$ across all comparisons in the YRI, $p < 0.05$ for the CEU) (Methods, Table 1). The slope on the autosomes for *R* (β_{AR}) is significantly larger than that for *G* (β_{AG}). The reverse inequality holds on the X chromosome, with both coefficients being larger on the X chromosome than on the autosomes. We note that $\beta_{XG} > \beta_{AG}$ is the primary finding of Hammer et al. (2010), where the interpretation is that linked selection on genes reduces diversity more rapidly on the X chromosome than on the autosomes.

The correlation between R and π/D

There are three possible explanations for the observed positive correlation between *R* and π/D: 1) recombination is itself mutagenic (Arbeithuber et al. 2015), 2) GC-biased gene conversion (gBGC) may influence allele frequencies in some way (Marias 2003) or 3) selection at noncoding (e.g. regulatory) sites are having a significant impact on local genetic diversity (recall that we only consider the distance from genes in the analysis above). Below we examine the evidence for each.

The local recombination rate of de novo mutations.

A direct measure of local mutation rate can be obtained by examining the distribution of *de novo* mutations identified via whole genome sequencing of families. We obtained two large genomic datasets consisting of 4,917 (Kong et al. 2012) and 11,010 (Francioli et al. 2015) such mutations, which we refer to as the Decode and GoNL datasets, respectively. To mirror our estimates of diversity, we masked repeats (Methods) and were left with 1,467 Decode and 2,166 GoNL mutations in the nonrepetitive

portions of the genome. We then tested if the recombination rate at sites with *de novo* mutations exceeded that of the genomic average.

In order to control for the effect of sequence context on mutation rate (for example enrichment of mutation rates at CpG sites, which in turn are generally enriched in recombination hotspots (Nachman 2001)) we generated our null distribution by extracting genome-wide sites that matched the trinucleotide distribution of the *de novo* mutations, yielding 67,394,212 and 60,392,816 sites in our null distributions for the Decode dataset and GoNL dataset, respectively.

When considering nonrepetitive regions, the Decode *de novo* dataset had a mean local recombination rate of 1.48 cM/Mb (SD: 4.90), while its matching null set had a mean of 1.47 cM/Mb (SD: 5.06). The GoNL *de novo* dataset had a mean local recombination rate of 1.57 cM/Mb (SD: 5.53), while its matching null set had a mean of 1.47 cM/Mb (SD: 5.06). Pooling the datasets yielded a mean *de novo* local recombination rate of 1.53 cM/Mb (SD: 5.06), which was not significantly different than the pooled null sets (p=0.48, 2-tailed Welch t-test). Repeating this procedure on the whole genome, including repetitive sequence but limited to the range of the HapMap genetic map, yielded a mean local recombination rate of *de novo* mutations of 1.54 cM/Mb (SD: 5.56) (1.51 cM/Mb SD: 5.38 for Decode, 1.56 cM/Mb SD: 5.63 for GoNL). The pooled trinucleotide-matching null set for Decode (n=227,640,296) and GoNL (n=206,352,696) had a mean local recombination rate of 1.37 cM/Mb (SD=4.72), which is significantly different from the combined set of 15,927 (4,917 Decode and 11,010 GoNL) *de novo* mutations (p=0.00013, 2-tailed Welch t-test). Thus, the expected local recombination rate of *de novo* mutations is significantly greater than that of the genomic average suggesting that recombination is associated with mutagenesis, consistent with previous works (Pratto et al. 2014; Arbeithuber et al. 2015; Francioli et al. 2015). However, when constrained to the nonrepetitive portions of the genome (as was used when examining the relationships with R and G above) there is no longer significant support for this effect.

GC-Biased gene conversion

Another cellular process that may induce a positive correlation between diversity and *R* (**Figure 2**) is gBGC. With gBGC, heterozygous genotypes near recombination-initiated double-strand breaks are preferentially converted from weakly bonded (W: A,T) to strongly bonded (S: G,C) states (Strathern et al. 1995; Marias 2003). This GC-biased DNA repair is posited to create a fixation bias whose effects have been shown in the AFS for the populations considered here (Katzman et al. 2011). As gBGC induces fixation biases for S→W and W→S mutations, we performed our heatmap and regression analyses using only S→S and W→W transversions, which are gBGC-independent. Visualizations of the joint effects of *R* and *G* on π/D under this condition largely recapitulate our original findings (**Figure 3**, Supplemental Figure S5), though the exclusion of so many variable sites increases the variance in our estimate of π/D. We repeated our IRLS regression analysis, and again found that all slope coefficients were significantly positive in both populations on both the X chromosome and the autosomes ($p < 0.001$), and that $\beta_{XR} > \beta_{AR} > \beta_{AG}$ in the YRI ($p < 0.001$ and in the CEU $p < 0.02$, Supplementary Table 1). Curiously, in both the YRI and the CEU we could not conclude that $\beta_{XG} > \beta_{XR}$. This may be due to reduced power as a result of considering ~1/6 the number of variable sites. We also note that our consideration of only S→S and W→W transversions serves as an additional control on the mutagenic effects of recombination, which is believed to be restricted to transitions (Arbeithuber et al. 2015). Overall we conclude that neither gBGC nor the mutagenicity of recombination are driving the trends seen in **Figure 1**.

The linkage to phastCons elements

If selection is driving the trend between π/D and *R*, it follows that this trend must largely stem from selection at nongenic sites. To evaluate this hypothesis we examined 725,430 phylogenetically conserved *phastCons* elements inferred over a primate phylogeny. Primate *phastCons* elements are short (mean 145

bp, SD 141 bp) genomic regions that have a marked lack of substitutions over the given phylogeny. Even though only ~27% of these elements overlap coding regions, they have long-term selective effects. We first computed the (minimum) genetic distance of each of our 10kb loci to conserved exonic (*phastCons*) elements (CEEs) and to conserved nonexonic (*phastCons*) elements (CNEs). To evaluate the effects of linkage to these elements we removed loci that overlapped either CNEs or CEEs, leaving 58,314 and 5,365 loci on the autosomes and X chromosome, respectively. As *phastCons* elements are so numerous, nearly all of the remaining loci were extremely close to at least one conserved element (see McVicker et al. 2009), with loci on the autosomes and the X chromosome having a mean minimum distance of 0.0067 cM (SD: 0.0261) and 0.0093 cM (SD: 0.0264), respectively. We then used IRLS regression to model the relative effects of linkage using the minimum genetic distances to CEE elements (D_{CEE}) and to CNE elements (D_{CNE}). We modeled $\pi/D \sim D_{CEE} + D_{CNE}$ separately for the X chromosome and the autosomes. Consistent with previous works (Hernandez et al. 2011, Halligan et al. 2013), π/D is positively correlated with both D_{CEE} and D_{CNE} ($p<0.001$, nonparametric bootstrap). Furthermore, our β-coefficient estimates are large for both the autosomes ($\beta_{\mathrm{DCNE}} = 0.057$, 95CI: 0.048-0.066 and $\beta_{\mathrm{DCEE}} = 0.057$, 95CI: 0.052-0.065) and the X chromosome ($\beta_{\mathrm{DCNE}} = 0.128$, 95CI: 0.116-0.154 and $\beta_{\mathrm{DCEE}} = 0.175$, 95CI:0.119-0.200). As a reference, when constrained to just nongenic loci that also do not overlap *phastCons* elements, modeling $\pi/D \sim G + R$ yields β-coefficients of $\beta_G = 0.041$, $\beta_R = 0.069$ and $\beta_G = 0.192$, $\beta_R = 0.079$ for the autosomes and X chromosome, respectively. β_{DCNE} and β_{DCEE} were not significantly different from each other within the X chromosome ($p=0.37$, nonparametric bootstrap) or within the autosomes ($p=0.81$, nonparametric bootstrap), suggesting that linkage to the nearest CNE and CEE have approximately equally impacts on diversity, though CNE are approximately three times more common than CEE

On the inference of neutralomes

Modeling linkage to conserved elements

In order to quantify the local effects of the selective forces described above, we used the approach of Halligan et al (using their Model C) to model the relative effective population size (ν, which is a prediction of π/D) for each locus based on the combined effects of all putatively selective *phastCons* bases within 1 cM of that locus. With our best-fitting model many loci have expected diversity levels that are much lower than the inferred neutral rate, with the X chromosome having a mean ν of 0.85 (SD: 0.13) and with the autosomes having a mean ν of 0.89 (SD: 0.12). Evaluations of the overall distribution of ν show more selective constraint on the X chromosome than the autosomes (**Figure 4**). This is consistent with previous findings (McVicker et al. 2009; Hammer et al. 2010; Gottipati et al. 2011; Veeramah et al. 2014).

Inferring sites unaffected by selection

We developed an approach to determine a cutoff for the minimum value for ν that removes the effects seen in **Figure 1**. Specifically, at this cutoff we lose all statistically significant positive correlations between any pair of choices among π/D, and *R*, *G* and ν. This can be framed as an approach to the problem of finding so-called "neutral" sequence, i.e. those loci where the distribution of allele frequencies are dictated solely by the rate of mutation and genetic drift and not by any direct or indirect (i.e. linked) effects of selection. We term the totality of all neutral loci in a genome the neutralome. We note that a lack of correlation between *R*, *G*, (and even ν) and π/D is a necessary condition of a sequence to be neutral.

Using statistical tests of both linear and of monotonic relationships we find that thresholding on $\nu \geq 99\%$ on the X chromosome and $\nu \geq 99.9\%$ on the autosomes is sufficient to remove any significant correlations between diversity, *R*, *G*, and ν. These results persist both when the autosomes and the X chromosome are evaluated separately, and when they are pooled. Pooling increases our power to detect correlations while mitigating concerns about the smaller sample size of the X chromosome. Visual

inspection of relationship between π/D and R further supports our assertion that the neutralome removes not just the significance, but also the trends seen in **Figure 1** (**Figure 5**). The neutralome's size, by genomic standards, is quite small, with only 3,606 and 787 10kb loci appearing neutral on the autosomes and X chromosome, respectively.

The AFS and estimates of Θ in different genomic regions.

To evaluate the properties of our neutralome we compared it against regions predicted to be neutral by Neutral Region Explorer (NRE) (Arbiza et al. 2012). NRE finds "neutral" regions parametrically, allowing the user to select specific criteria (e.g. thresholds on the minimum genetic distance to genes, local recombination rate, and background selection coefficients (McVicker et al. 2009)) that may (or may not) lead to regions unaffected by selection. We downloaded regions identified as neutral by NRE using all parameters set to their defaults, save our removal of the filter on the local recombination rate. We took the 10kb loci from the YRI that overlapped the genomic coordinates identified as neutral by NRE as a candidate set of "neutral" NRE loci. NRE yielded more loci than our neutralome (10,094 vs 4,393), with an overlap of 764 regions. For the autosomes, adding an additional filter on the background selection coefficient of McVicker et al. (2009), with a setting of 98% neutral (BG98), yielded a dataset comparable in size to the neutralome (n=3,950). We then repeated out analysis of these NRE-BG98 loci. The default settings of 95% neutral yielded a comparable subset of loci on the X chromosome (n=170).

For all sets of loci we computed two estimates of the population mutation rate θ, θ_W, θ_π, dividing each by D to correct for mutation rate heterogeneity, as well as Tajima's D (Tajima, 1989), Fu and Li's D (Fu and Li, 1993), and Thomson's estimator of the time to the most recent common ancestor (TMRCA) (Hudson, 2007). The means of each summary statistic were compared between loci exclusive to our neutralome versus loci exclusive to NRE All significance tests were 1-tailed Welch t-tests, with the expectation being that our neutralome would have higher values of all summary parameters.

Our neutralome has significantly higher mean θ/D (A: θ_π 0.083; X: θ_π 0.072) than either the default (A: θ_π 0.074, $p < 2.2e\text{-}16$; X: θ_π 0.066, $p < 0.05$) or the NRE-BG98 subset (A: θ_π 0.077, $p < 1e\text{-}10$; X: θ_π 0.060, $p < 0.07$) (Supplemental Tables 3, 4, **Figure 6**) for both the autosomes and the X chromosome, indicating that NRE regions are more affected by linked selection. The neutralome also has significantly higher Tajima's *D* and Fu and Li's *D* (**Figure 7**), indicating an excess of rare variants in the NRE loci. While the significance of some of these effects disappear for NRE,BG98 loci, there still remains a significant effect for Tajima's *D* on the X chromosome even under this more stringent criterion (n=21). Lastly, we see a striking reduction in TMRCA for both sets of NRE loci, though this likely stems from positive correlations between TMRCA and θ.

To assess if the neutralome impacts demographic inference we fit the AFS to a 2-epoch instantaneous growth model using ∂a∂I (Gutenkunst et al. 2009) in the YRI. We estimated the ancestral population size (Na), the growth rate (*n*), and the time of growth (*T*) assuming a generation time of 25 years and mutation rate of 2.5×10^{-8} (Nachman and Crowell, 2000). These parameters were estimated using loci that are exclusive to NRE, loci that are exclusive to the neutralome (and not NRE), and for a reference we also computed the AFS using 4-fold degenerate sites (table 2). Consistent with our estimates of θ, Na was significantly higher in the neutralome than in regions found by NRE and in 4-fold degenerate sites (table 2, bootstrapped $p \leq 0.001$ across all comparisons). And, consistent with our analysis of Fu and Li's D, *n* in the neutralome was significantly less across the same comparisons (table 2, bootstrapped $p < 0.001$ across all comparisons), indicating an excess of rare variants in the NRE loci and in 4-fold degenerate sites. Further, the time of growth *T* varied significantly across our choices of "neutral" samples, with the neutralome (~161 thousand years ago, kya) yielding older growth estimates than the NRE loci (~125 kya) and 4-fold sites (~94 kya). While the *p*-values for *T* were generally larger than for the other parameters, they still obtained statistical significance across all comparisons (bootstrapped $p < 0.05$ across all comparisons, table 2).

DISCUSSION

The pervasive effects of linked selection

Using high coverage whole genomes from two human populations we show that even if only ~5-15% of sites in the human genome are directly targeted by selection (Chinwalla et al. 2002, Cooper et al. 2005, Siepel et al. 2005, Meader et al. 2010, Ponting and Hardison, 2011, Rands et al. 2014), the indirect action of selection at linked sites causes genetic diversity in up to 99% of the genome to be demonstrably reduced from neutral expectations. On genomic scales, the effects of linked selection up to this point have solely been framed either with respect to how linked a particular locus is to the nearest genic source of selection (Hammer et al. 2010, Gottipati et al. 2011, Arbiza et al. 2014), or with correlations between diversity, gene-density and rates of recombination (Cai et al. 2009, Lohmueller et al. 2011, McGaugh et al. 2012). We examined the dependence of π/D against two genomic measures sensitive to selection at linked sites, the minimum genetic distance to genes and the local rate of recombination, and show that genetic diversity is influenced by both measures at fine scales throughout the genome (**Figure 1**). While both *R* and *G* are positively correlated with diversity, the effect of *R* is far greater than that of *G* on the autosomes (**Table 1**). This effect is neither driven by GC-biased gene conversion nor the mutagenicity of recombination (**Figure 3**). Instead we find that these patterns are primarily driven by linked selection at nongenic sites, an argument that is bolstered when we consider the sizable effect of *R* on diversity in loci that are generally unlinked from all genic sites of selection (the last column of **Figure 1**).

Our hypothesis of the effect of linked selection at both genic and non-genic loci is consistent with several other aspects of our findings. First, linked selection may have sizable impacts on diversity, while it has at most a modest effect on patterns of divergence (Birky and Walsh, 1988, but see McVicker et al. 2009). Consistent with this both π and *D* are correlated with *R* and *G*, and the effect is much stronger for π than *D* (Supplemental Figures S3, S4). Examination of these patterns on the X chromosome versus the

autosomes also supports our linked-selection hypothesis. As the X chromosome is exposed to selection in males, beneficial recessive alleles are much more likely to be driven to fixation, while alleles with equivalent selection coefficients are more likely be lost to drift on the autosomes (Maynard Smith and Haigh, 1974, Charlesworth, 1996). Veeramah et al. (2014) estimated that 46-51% of nonsynonymous mutations are driven to fixation by positive selection on the X chromosome, compared to 4-24% on the autosomes. Given this vast disparity in the rates of genic adaptive substitution between the X chromosome and the autosomes in humans, and the hitchhiking that results from this, it is perhaps unsurprising that the effect size for *G* on the X chromosome (β_{XG}) is greater than that of the autosomes (β_{AG}) (**Table 1**), consistent with previous works (Hammer et al. 2010, Gottipati et al. 2011, Arbiza et al. 2014). In addition, for linked selection that includes nongenic sites, the effect for *R* on the X chromosome (β_{XR}) is larger than that of the autosomes (β_{AR}) (**Table 1**), suggesting that this faster X effect may be acting on regulatory elements in addition to genic elements, though perhaps to a lesser degree as $\beta_{XG} > \beta_{XR}$. Attributing $\beta_{XR} > \beta_{AR}$ to other nonselective causes, however, is more difficult, and perhaps requires a mechanistic sex-biased explanation consistent with the X chromosome spending 2/3 of its time in females.

In considering the role that nongenic selective sites may have on population genetic diversity we examined the effects of being linked to phylogenetically conserved *phastCons* elements inferred in primates. Using the same metrics as with *G*, we show that the effect size for being linked to the nearest conserved nonexonic elements (CNE) is as great as it is for conserved exonic elements (CEE) on both the X chromosome and the autosomes. This not only supports the hypothesis that linked selection to conserved (putatively regulatory) sites is as important as exonic sites, but as CNEs outnumber CEEs nearly 3:1, it may suggest that the cumulative effects of being linked to regulatory sequence may be greater than the cumulative effects for being linked to protein-coding sequence.

The human neutralome

While our findings are consistent with the action of selection at linked sites, the extent that the linked selection to *phastCons* elements drives the patterns seen in **Figure 1** remains to be seen. Finding loci unlinked to every *phastCons* element is, however, infeasible as they are both numerous and dispersed. Their small size and distributed nature necessitates a more comprehensive measures on the amounts of linked selection. Linkage to a single *phastCons* element may, for example, have little effect. To this end we applied the approach of Halligan et al. (2013) to infer the relative effective population size (ν) for each of our 10kb loci. The genomic distribution of ν shows that while ~70% genomic bases are somewhat linked to sites of selection (ν = 80-99%), only ~2% are completely decoupled from sites of selection (ν > 99.9% on the autosomes) (**Figure 4**).

The issues connected to the use of ν to determine where the trends displayed in **Figure 1** reverse are delicate. Notably, *phastCons* elements are unlikely to be the sole targets of selection in the genome. In addition, genic sites that do not intersect *phastCons* elements are not included in our estimate of ν. A simpler, and testable, approach is to assume that selective sites occur in a background that contains some appreciable density of *phastCons* elements. With this assumption, it follows that if ν is sufficiently large—that is, we are generally far from the linked effects of *phastCons* elements—then perhaps the trends in **Figure 1** will no longer be apparent. Thresholding solely on ν is sufficient to remove both gradients seen in **Figure 1**, as well as any significant correlations with ν itself (**Figure 5**, Supplemental Figure S6). The high thresholds on ν lead to a diminutive human neutralome and demonstrate that linked selection, rather than some other mechanistic force, is the primary contributor to the correlation between π/D and G, R.

The neutralome versus other definitions of neutrality

Demographic inferences are based on the assumption that the sequences considered are "neutral." We show that regions identified as neutral by NRE—regions that consider both linkage to genes and background selection in their definitions—lead to statistically significantly different values for the summary statistics than that of the neutralome. The neutralome has higher effective population size than the loci identified by NRE (**Figure 6**, Supplemental Tables 3, 4), further, it has fewer low-frequency polymorphisms, as measured by Tajima's *D* (Supplemental Tables 3, 4) and Fu and Li's *D* (**Figure 7)**. This suggests that linked selection is not only reducing θ in the NRE loci, but it is altering the distribution of the AFS (Tachida 2000; Williamson and Orive 2002; Nicolaisen and Desai 2012, 2013). While incorrect estimates of θ may only bias parameter estimates by a constant factor, changes to the AFS may fundamentally alter the demographic inference. To assess this we fit a simple 2-epoch instantaneous growth model using ∂a∂I (Gutenkunst et al. 2009) considering sites in the neutralome, sites found by NRE, as well as in 4-fold degenerate sites. All three demographic parameters varied considerably across these definitions of "neutral" sites, with the use of regions with greater selective impacts (4-fold degenerate sites and loci found by NRE) inferring a smaller effective population size, as well as larger and more recent population growth. Taken together this suggests the linked selection biases demographic inference in regions either classically considered to be neutral (4-fold sites), as well as sites that have been carefully chosen to minimize the effects of linked selection (NRE). While skews in the allele frequency spectrum, and the demographic biases that result, may seem to have a limited scope, demographic models serve as powerful null models for inferences on selective processes (e.g., Kieghtley and Eyre-Walker 2012; Singh et al. 2013; Veeramah et al. 2014; Hsieh et al. 2016; Uricchio et al. 2016). Thus, it follows that biased null demographic models may induce bias in the testing of alternative models of selection, implying that mis-specified demographic models may impact inferences on both neutral and selective processes.

CONCLUSIONSs

Intrinsic molecular forces, such as rates of mutation and recombination, and extrinsic forces, both demographic (e.g. migration and drift) and selective (e.g., linked selection), together shape patterns of diversity throughout the genome. Here we show that linked selection reduces diversity throughout the vast majority of sites in the human genome. This finding has several consequences. The first is with respect to predictions on the number of selective bases in the genome. We find that conditioning on linkage to *phastCons* elements is sufficient to find loci with no significant evidence of linked selection. Thus, if sites other than *phastCons* elements are reducing diversity, it follows that these sites likely occur in *phastCons*-rich genomic regions (e.g., Cheng et al. 2014). These genome-wide reductions in diversity also likely impact direct inferences, such as neutral estimates of effective population size and the effective population size of the X chromosome compared to the autosomes. Future work in genomics may need to account for these linked selective effects, which in turn necessitates the development of better null models—models that permit variable Θ and variation in the AFS from background selection—that are best-equipped to separate true signals of natural selection from that of the genomic background.

METHODS

Samples and Locus Preparation. We used a total of 18 high coverage whole genomes, 9 West African samples (YRI) and 9 European samples (CEU), made publicly available by Complete Genomics (Drmanac et al. 2010). We computed nucleotide diversity (π) across the genome in nonoverlapping 10kb windows of the hg19 genome. Divergence was computed from the 46-way Multiz vertebrate alignments made available from the UCSC genome browser website (Kent et al. 2002). Using the rhesus macaque to polarize sites, we computed the average divergence between each population and the ancestor of humans and orangutans. Both π and D were computed using a mixture of male and female samples. To perform

such calling on the X chromosome we utilized the ploidy information imbedded in the .tsv files for all samples. Similar to our *de novo* analysis (below), sites that fell in microsatellites, simple repeats, repetitive elements, segmental duplications, self-chain regions, as well as regions identified as copy number, structural variants or as uncallable by Complete Genomics, were removed from our analysis. After masking, 10kb windows with ≤1kb of callable sequence in either population or that were outside of our genetic map were also removed, giving a total sample size of 244,661 autosomal loci and 12,491 X chromosome loci.

We used the population averaged (YRI+CEU) genetic map from HapMap (Frazer et al. 2007), and scaled the X chromosome by ⅔ to compute all genetic distances. The UCSC known genes track (downloaded on 12/04/2015) (Hsu et al. 2006), which includes both protein coding and RNA genes, was analyzed using a custom perl script. For each locus, we computed in centimorgans (cM), the genetic length *R* and the minimum distance to genes *G*.

Regression Analysis.

Linear regression has many assumptions that are frequently violated in genomic analyses. In particular, outliers can have sizable impacts on parameter estimates, with such outliers in genomics arising, for example, from unmasked copy-number variants or structural variants. While outlier-detection and removal is one possible method, a more comprehensive approach is the use of "robust" regression techniques that reduce the impact of influential observations. Using the R statistical package MASS, we performed iterated reweighted least squares (IRLS) regression (i.e. "robust" regression) using the rlm command set to the default parameters. IRLS reduces the impact of influential observations by performing a linear regression using an iterative process until both the weights and the regression parameter estimates converge. In particular, the default setting with rlm uses Huber weighting in which the initial set of weights are 1. As a function of the residual *e*, weights *w* are defined by

$$w(e) = \begin{cases} 1 \ for\ |e| \leq k \\ \frac{k}{|e|} for\ |e| > k \end{cases}$$

reducing the weights for outliers (residual $e > k$) far from the inferred hyperplane. The cutoff k is determined using convex optimization. Note that IRLS produces unbiased coefficient estimates in the presence of either autocorrelation (due to linkage disequilibrium) or heteroscedasticity, both of which are present here. Because estimates of the variance of the slope parameter estimates may be biased downwards, we used a 1000-iteration nonparametric bootstrap to assess the significance of the slope signs and slope differences between the X chromosome and the autosomes. The bootstrapped *p*-values are the fraction of bootstrapped slopes with coefficients greater than zero (for slope signs) and the fraction of bootstrap samples where one fixed slope coefficient was larger than the other (for slope differences).

Local recombination and de novo mutation.

The Decode dataset was obtained from the supplemental information provided by Kong et al. (2012), which provides the location of 4,933 autosomal *de novo* mutations in the hg18 reference genome. The Decode dataset was augmented with a similar, but larger (n=11,020), list of autosomal *de novo* SNPs called from the Genome of the Netherlands (GoNL) project (Francioli et al. 2015) annotated in the hg19 genome. The GoNL dataset was mapped to the hg18 genome using UCSC's liftOver utility (Hinrichs et al. 2006), putting all mutations in the same coordinate system. As we are primarily concerned with properties of recombination in the nonrepetitive portions of the genome, we restricted all mutations to those that fall within the HapMap genetic map (Frazer et al. 2007) and the genetic map of Kong et al. (2010). We further removed mutations that fell in our mask of segmental duplications (Bailey et al. 2002), self-chain sequence, copy number variants (CNVs) (MacDonald et al. 2013), Numts (Lascaro et al. 2008), microsatellites, simple repeats or repeat elements, with the uncited annotations coming from the hg18

reference genome tables from UCSC. This left 1,467 SNPs in the Decode dataset, and 2,166 in the GoNL dataset that fell in the nonrepetitive portion of the genome available and thus were available for analysis.

After masking the genome using the same set of filters as above, we computed the distribution of all 32 possible trinucleotides (assuming strand asymmetry) in the unrepetitive parts of the genome as well as of our *de novo* datasets. Using reservoir sampling, we created a maximally-sized exactly-matching distribution of null sites in the genome. This procedure was applied separately to match the GoNL dataset and the Decode dataset, yielding 60,392,816 and 67,394,212 sites for use in our null distributions, respectively. Using bedtools (Quinlan and Hall, 2010), we computed the local recombination rate of both our *de novo* mutations and our null distributions.

***phastCons* element analysis**

After querying the UCSC genome browser for *phastCons* elements inferred in primates we applied Ensembl's (version 72) perl API (Yates et al. 2016) to find exons from canonical protein-coding genes. We divided *phastCons* conserved elements into two categories; 196,044 elements that intersected protein-coding exons (conserved exonic elements, or CEEs), and 529,386 elements that did not (conserved nonexonic elements, or CNEs). Using our 10kb loci, including those that overlap genes, we computed the minimum genetic distance to CEE and CNE elements using the HapMap genetic map (Frazer et al. 2007).

Estimating effective population size

We apply "Model C" in Halligan et al. (2013), a nonlinear regression model for $\log(\pi/D)$, that considers both coding and noncoding elements. In Model C, diversity at linked sites decays exponentially near selective elements, with the predicted diversity at a neutral site being a product over all linked selected sites. The diversity at a single linked site is approximated by:

$$\frac{\pi}{D} \sim \exp\left(\log(p_1) - p_2 \sum e^{-\frac{x_i}{p_3}} - \ p_4 \sum e^{-\frac{x_i}{p_5}}\right)$$

where p_1 is the neutral or unreduced level of diversity, p_2 is the reduction in diversity observed at a single CEE site, p_3 is the rate of decay around this CEE site, p_4 is the reduction in diversity at a single CNE site, p_5 is the rate of decay around a single CNE site, and x_i is the distance in morgans $(M)^{-8}$ from a neutral site to the *i*th selective site. We assume that the rate of recombination is constant within an element (a reasonable assumption as *phastCons* elements have a mean length of 145 bp), which is several times shorter than the resolution of the HapMap genetic map.). The computation time was reduced by summing over selective elements rather than selective bases.

Using Model C , we estimated the expected diversity at the center of each 10kb locus based on the extent of that locus's linkage to all CEE and CNE elements within 1 cM. Using Nelder-Mead optimization under a sum of squares criterion we found parameters to maximize model fit. To take into account issues of nonuniqueness in nonlinear least-squares regression we employed a 2-tiered sampling approach to generate estimates of diversity across the genome. For the autosomes, we randomly sampled 5% of the loci and chose 5 random values for the starting points for optimization. After repeating this process 100,000 times we chose the top 10,000 distinct stopping points found by the optimization and treated these as starting positions for optimization for the entire autosomal dataset. As the X chromosome is roughly 5% of the size of the autosomes we randomly drew 100,000 starting points for optimization and chose the parameters that yielded the best model fit for the entire X chromosome. Both the autosomes and the X chromosome yielded parameter estimates of p_1-p_5 that varied substantially, yet yielded exceedingly similar R^2 values (Supplemental Table 2). We chose the ~~top~~ 100 best fitting parameters (Supplemental Table 2), computed the minimum and the variance in our relative diversity estimates, and noted that the vast majority of their predicted diversity values are exceedingly similar. Further, when contrasting the minimum expected diversity relative to the variance in this prediction across the top 100

best fitting models, we observed that the models largely agreed on which loci should be called neutral, yet varied sometimes substantially when at least one of the model estimates predicted low diversity. Consequently, disparate parameters yield largely consistent predictions of diversity for loci that are nearly neutral. We also repeated our experiments using the union, as opposed to the intersection, of exons and conserved elements that intersected exons to form our CEE class and arrived at model fits that were indistinguishable from our current results (results not shown).

On the inference of "neutral" sequence

phastCons elements make up a substantial subset of sitess that are under diversity-reducing selection in the human genome. While Model C can in principle be extended to other classes of selective elements, this extension comes at considerable computational expense. Focusing on the goal of determining a set of neutral regions useful for demographic inference among other needs for genomic null distributions, we establish candidate sets of effectively neutral loci. We begin by evaluating correlations between π/D, and R, G, or ν conditioned on thresholds on ν (the inferred relative effective population size). Interpreting these correlations must be done with care as neutrality is evidenced by a lack of a statistically significant difference from zero. We must be concerned whether this lack of significance can be attributed to a lack of power (e.g. stemming from too small a sample size), poor model choice (e.g. using a linear model when there are significant non-linear effects), or issues resulting from multiple testing. We are also concerned with whether the neutralome we find is less "neutral" on the X chromosome simply owing to reduced power stemming from its smaller size. To address these concerns we use both our original IRLS regression framework and an extension that pools the samples for the X chromosome and the autosomes. To a monotonic relationships with π/D., we use Kendall's rank correlation test separately for the X chromosome. Lastly we limit the number of candidate sets of neutral loci that we evaluate to limit multiple testing issues. As our tests do not account for linkage disequilibrium, the reported *p*-values are

susceptible to downward bias. Because we are looking for a lack of significance, this makes our overall approach conservative to the question of finding neutral loci.

In the preceding sections we modeled $\pi/D \sim G + R$ separately for the X chromosome and the autosomes and then tested for slope-signs and slope-differences. For the sake of completeness, let's consider an alternative approach for testing slope differences with respect to the (marginal) effect of *R*. We can model the X chromosome and the autosomes jointly in a single regression: $\pi/D \sim R + X + RX$ with an indicator variable *X* indicating whether the locus is on the X chromosome. This produces coefficient estimates (β) of the form: $\beta_0 + \beta_1 R + \beta_2 X + \beta_3 RX$. For autosomal loci (X=0), this model produces an estimate of the intercept (β_0) as well as the coefficient for *R* (β_1), and for loci on the X chromosome (X=1), after collecting like terms, the model estimates a different intercept ($\beta_0 + \beta_2$) and a different slope ($\beta_1 + \beta_3$). In this way, the significance of the slope difference and intercept difference can be ascertained from the significant difference from 0 of the β_2 and β_3 parameters, respectively. Thus testing for a slope-difference between the X chromosome and the autosomes is framed within a single regression. Further, upon finding for a lack of significance in the interaction term *RX,* the variances across the X chromosome and the autosomes for the slope can be pooled simply by omitting this term and the *X* term

Given the above, we found a neutralome using the following strategy. Using increasing thresholds on ν, we form candidate sets of neutral loci. We use these candidate sets to find the minimum ν with no significant correlations between pairs from π/D, *R*, *G*, or ν using both linear models that combine and separate sample sizes across the X chromosome and the autosomes (as per the preceding paragraph), as well as with nonparametric models to test for significant monotonicity. For a given value of ν, we used IRLS regression to model $\pi/D \sim R + G + \nu$ separately for the X chromosome and the autosomes, with the *p*-values from the IRLS regression being obtained from the coeftest function in the lmtest library in R using the default parameters. After testing an initial threshold of 0.98, increasing our threshold to $\nu > 0.99$ removed all significant effects for the X chromosome (with a minimum *p*-value of 0.14), yet significant

effects remained on the autosomes, with *p*-values of 0.02, 7e-8 and 2.2e-16 for the coefficients of *R*, *G*, and ν, respectively. For the autosomes, we found that the criterion of $\nu > 0.999$ was sufficient to remove any significant correlations (minimum *p*-value of 0.34). Next, we recapitulated our findings by modeling $\pi/D \sim R + G + \nu + X + RX + GX + X$, for loci with $\nu > 0.99$ on the X chromosome and $\nu > 0.999$ on the autosomes, and we only found significant effects for the intercepts. We then pooled variances across the X chromosome and the autosomes by modeling $\pi/D \sim R + G + \nu + X$, which again showed no significant effects for *R*, *G*, or ν with negative slope coefficient point estimates for *R* and ν and a *p*-value of 0.24 for *G*. To accommodate concerns of lack of statistical power, we employed the nonparametric Kendall's rank correlation test for (marginal) monotonic relationships by evaluating the *p*-value of the correlation coefficient (τ) using the cor.test function in R separately for both the X chromosome and the autosomes. Tests for correlations between π/D and *R*, *G*, or ν were not significant (minimum *p*-value across all six tests is 0.17). This yielded a neutralome of ~44 million bases of sequence consisting of 3,606 10kb loci on the autosomes and 787 10kb loci on the X chromosome.

Demographic inference

Using ∂a∂I (Gutenkunst et al. 2009), we fit folded allele frequency spectra under a simple 2-epoch instantaneous growth model in the YRI. This model was fit considering sites exclusive to the neutralome (and not found by NRE), sites exclusive to NRE (and not found in the neutralome), as well as with 4-fold degenerate sites. 4-fold degenerate sites were obtained from SnpEff (version 3.3) (Reumers et al. 2008) which used Ensembl (Version 72) to identify canonical protein-coding 4-fold degenerate sites. To remove repeats 4-fold degenerate sites were then masked using the same set of masks used to compute diversity and divergence. ∂a∂I was then used to infer three parameters: the ancestral population size (Na), the ratio of the contemporary to the ancestral population size (*n*), and the time *T* in generations of the population size change. Point estimates for these parameters were obtained using the 2-epoch model fitting described

in Veeramah et al. (2014). Briefly, a coarse two-dimensional grid in *n* and *T* were searched with *n* varying from 0.1 to 10, and *T* varying from 0 to 2.0, both in steps of 0.1. The best parameter set in this 2-dimensional space was then further refined using Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization. Confidence intervals in our parameter estimates were estimated using a 1000-iteration bootstrap.

DATA ACCESS

The code used to generate our estimates of ν across the genome, as well as the location of human neutralome are available at: http://hammerlab.biosci.arizona.edu/Neutralome/neutralome.html

ACKNOWLEDGMENTS

Support for this work was provided by the US National Institutes of Health to M.F.H. (R01_HG005226) and NSF Graduate Research grant (DGE-1143953) to A.E.W.

DISCLOSURE DECLARATION

None

REFERENCES

1. Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci* **112**: 2109-2114.
2. Arbiza L, Zhong E, Keinan A. 2012. NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics* **14**:1.
3. Arbiza L, Gottipati S, Siepel A, Keinan A. 2014. Contrasting X-linked and autosomal diversity across 14 human populations. *Am J Hum Genet* **94**:827-44.
4. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**: 1003-7.
5. Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*. **356**: 519-520.
6. Begun DJ, Whitley P. 2000. Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proc Natl Acad Sci 97:5960-5.
7. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321-5.
8. Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci* **85**: 6414-8.
9. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**: 783-96.
10. Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and

regulatory sites in humans. *PLoS Genet* **5**: e1000336.
11. Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289-303.
12. Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* **141**: 1619-32.
13. Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80**: 692-704.
14. Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, Euskirchen G. 2014. Principles of regulatory information conservation between mouse and human. Nature. 515: 371-5.
15. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*. **420**: 520-62.
16. Comeron JM. 2014. Background selection as baseline for nucleotide variation across the Drosophila genome. *PLoS Genet*. **10**: e1004434.
17. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**: 901-13.
18. Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol*. **13**: e1002112.
19. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. **327**: 78-81.
20. Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* **26**: 2097-108.
21. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, van Duijn CM, Swertz M, Wijmenga C, van Ommen G et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**: 822-6.
22. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-861.
23. Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693-709.
24. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A. 2011. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat Genet* **43**:741-3.
25. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. **5**: e1000695.
26. Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. *Mol Biol Evol* **28**: 2651-60.
27. Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet.* **5**;9:e1003995.
28. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD. 2010. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nat Genet* **42**: 830-831.
29. Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res*. **8**: 269-94.

30. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J. 2006. The UCSC genome browser database: update 2006. *Nucleic Acids Res*. **34**: D590-8.
31. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920-4.
32. Hsieh P, Veeramah KR, Lachance J, Tishkoff SA, Wall JD, Hammer MF, Gutenkunst RN. 2016. Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res* **26**: 279-90.
33. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics*. **22**: 1036-46.
34. Hudson RR. 2007. The variance of coalescent time estimates from DNA sequences. *J Mol Evol* **64**: 702.
35. Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics*. **120**: 831-40.
36. Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* **141**: 1605-17.
37. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science*. **317**: 915
38. Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol* **3**: 614-626.
39. Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J Mol Evol*.. **74**: 61-8.
40. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci*. **111**: 6131-8.
41. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res*. **12**: 996-1006.
42. Khachane AN, Harrison PM. 2009. Assessing the genomic evidence for conserved transcribed pseudogenes under selection. *BMC Genomics*. **10**: 435.
43. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**:1099-103.
44. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, igurdsson A, Jonasdottir A, Jonasdottir A et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471-475.
45. Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M. 2008. The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. *BMC Genomics* **9**: 1.
46. Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, et al.. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet* **7**: e1002326.
47. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2013.The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**: D986-92.
48. Marias, G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet* **19**: 330-338.

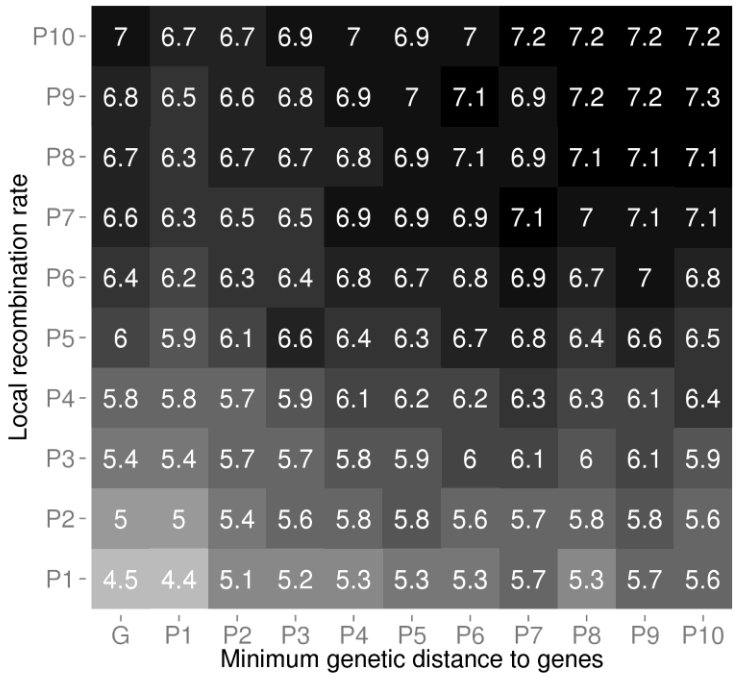
49. Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23-35.
50. McGaugh SE, Heil CS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MA. 2012. Recombination modulates how selection affects linked sites in Drosophila. *PLoS Biol*. **10**: e1001422.
51. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*. **304**: 581-4.
52. McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5**: e1000471.
53. Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20**: 1335-43.
54. Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics*. **156**: 297-304.
55. Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *TIG*. **17**: 481-5.
56. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, Reese AL, McAnally JR, Chen X, Kavalali ET, et al. 2016. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*. **351**: 271-5.
57. Nicolaisen LE, Desai MM. 2012. Distortions in genealogies due to purifying selection *Mol Biol Evol* **29**: 3589-3600.
58. Nicolaisen LE, Desai MM. 2013. Distortions in genealogies due to purifying selection and recombination. *Genetics*. **195**: 221-30.
59. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. **3**: e170.
60. Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genet Res* **67**: 159–174.
61. O'Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol* **27**: 1162-72.
62. Ponting CP, Hardison RC. 2011.What fraction of the human genome is functional? *Genome Res* **21**: 1769-76.
63. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A. 2013. Great ape genetic diversity and population history. *Nature*. **499**: 471-5.
64. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. *Science* **346**:1256442.
65. Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet*. **10**: e1004525.
66. Reumers J, Conde L, Medina I, Maurer-Stroh S, Van Durme J, Dopazo J, Rousseau F, Schymkowitz J. 2008. Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nuc Acid Res*. **36**: D825-9.
67. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-2.
68. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034-50.

69. Singh ND, Jensen JD, Clark AG, Aquadro CF. 2013. Inferences of demography and selection in an African population of Drosophila melanogaster. *Genetics*. **193**: 215-28.
70. Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet*. **2**: e148.
71. Strathern JN, Shafer BK, McGill CB. 1995. DNA synthesis errors associated with double-strand-break repair. *Genetics* **140**: 965-972.
72. Tachida, H. 2000. DNA evolution under weak selection. *Gene* **261**: 3-9.
73. Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-95.
74. Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**: 2325-7.
75. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, Sninsky JJ, Cargill M, Adams MD, Bustamante CD et al. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet* **5**: e1000592.
76. Uricchio LH, Zaitlen NA, Ye CJ, Witte JS, Hernandez RD. 2016. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res* **26**: 1-11.
77. Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. 2014. Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. *Mol Biol Evol* **31**: 2267-82.
78. Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci **103**: 135–140.
79. Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol Biol Evol* **19**: 1376-84.
80. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet*. **3**: e90.
81. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG. 2016. Ensembl 2016. *Nucleic Acids Res*. **44**: D710-6.

FIGURES

Figure 1. Heatmap of the median π/D (x 100) in YRI. The x-axis shows the minimum genetic distance to genes and the y-axis the local recombination rate. Darker cells correspond to higher π/D. Each cell corresponds to a pair of percentiles (P) in the distance to genes and the local recombination rate, with column G corresponding to loci in genes, and the remaining columns being outside of genes. a) Autosomes. b) X chromosome

a


P10 7 6.7 6.7 6.9 7 6.9 7 7.2 7.2 7.2 7.2
P9 6.8 6.5 6.6 6.8 6.9 7 7.1 6.9 7.2 7.2 7.3
P8 6.7 6.3 6.7 6.7 6.8 6.9 7.1 6.9 7.1 7.1 7.1
P7 6.6 6.3 6.5 6.5 6.9 6.9 6.9 7.1 7 7.1 7.1
P6 6.4 6.2 6.3 6.4 6.8 6.7 6.8 6.9 6.7 7 6.8
P5 6 5.9 6.1 6.6 6.4 6.3 6.7 6.8 6.4 6.6 6.5
P4 5.8 5.8 5.7 5.9 6.1 6.2 6.2 6.3 6.3 6.1 6.4
P3 5.4 5.4 5.7 5.7 5.8 5.9 6 6.1 6 6.1 5.9
P2 5 5 5.4 5.6 5.8 5.8 5.6 5.7 5.8 5.8 5.6
P1 4.5 4.4 5.1 5.2 5.3 5.3 5.3 5.7 5.3 5.7 5.6
Local recombination rate
G P1 P2 P3 P4 P5 P6 P7 P8 P9 P10
Minimum genetic distance to genes

b

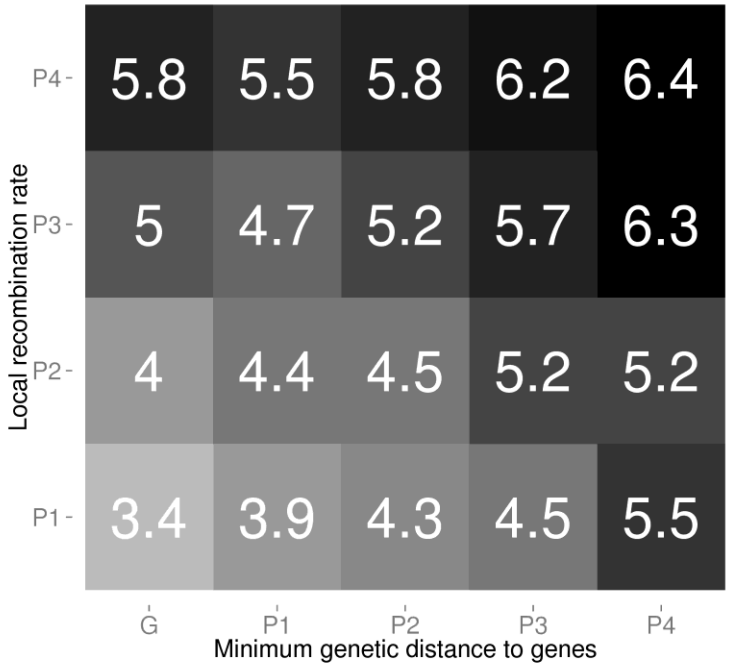

P4 5.8 5.5 5.8 6.2 6.4
P3 5 4.7 5.2 5.7 6.3
P2 4 4.4 4.5 5.2 5.2
P1 3.4 3.9 4.3 4.5 5.5
Local recombination rate
G P1 P2 P3 P4
Minimum genetic distance to genes

Figure 2. The role of base composition on different mutation types. GC-biased gene conversion (gBGC) increases the fixation probability of strongly bonded base pairs (S, G or C) over weakly bonded base pairs (W, T or A). Thus, gBGC can increase diversity (upwards arrow, blue) or decrease diversity (downwards arrow, red) for transitions (I) and transversions (V) alike. Similarly, recombination-induced mutagenesis is almost entirely exclusive to CpG mutations (which are S→W transitions) and may be restricted to transitions (Arbeithuberet al. 2015). To avoid both phenomena, we repeated our experiments using the beige class of S→S and W→W transversions (beige).

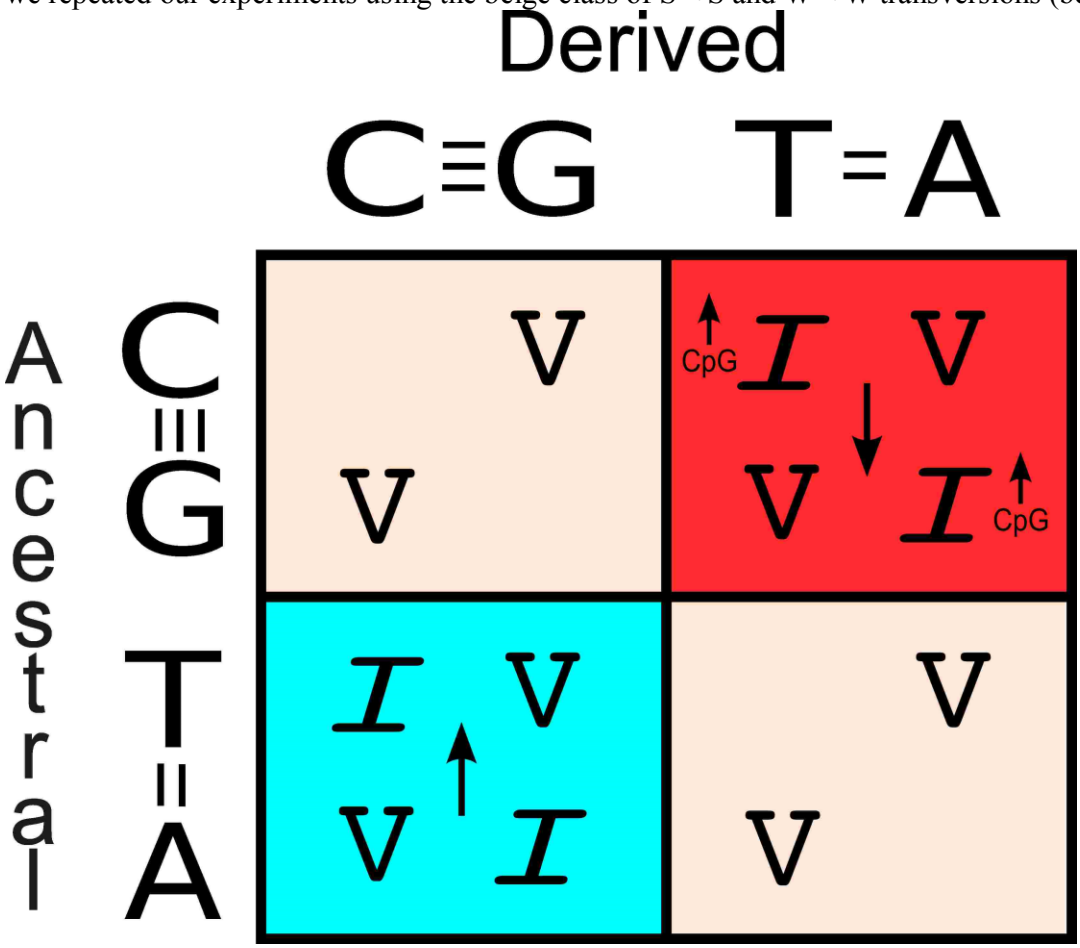


Figure 3. Heatmap of the median π/D (x 100) in YRI considering only W→W and S→S transversions. The x-axis shows the minimum genetic distance to genes and the y-axis the local recombination rate. Darker cells corresponding to higher π/D. Each cell corresponds to a pair of percentiles (P) in the distance to genes and the local recombination rate, with column G corresponding to loci in genes, and the remaining columns being outside of genes. A) Autosomes. B) X chromosome

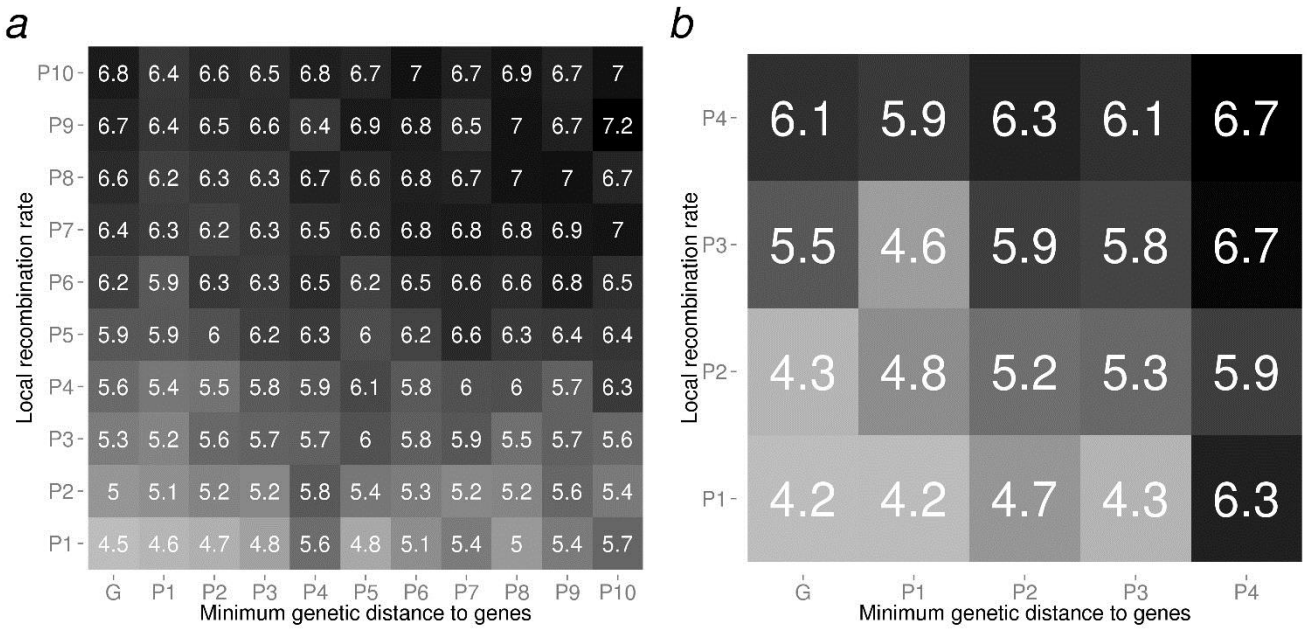


Figure 4. The empirical cumulative distribution function (ECDF) of the predicted nucleotide diversity relative to a neutral rate of 1. Predictions are based off of nonlinear least squares fitting of observed diversity levels to linkage to *phastCons* elements. In general, the X chromosome (gray) shows more constraint than the autosomes (black), with many loci having diversity close to the neutral rate (~0.9), but few loci having diversity levels at the neutral rate (1.0).

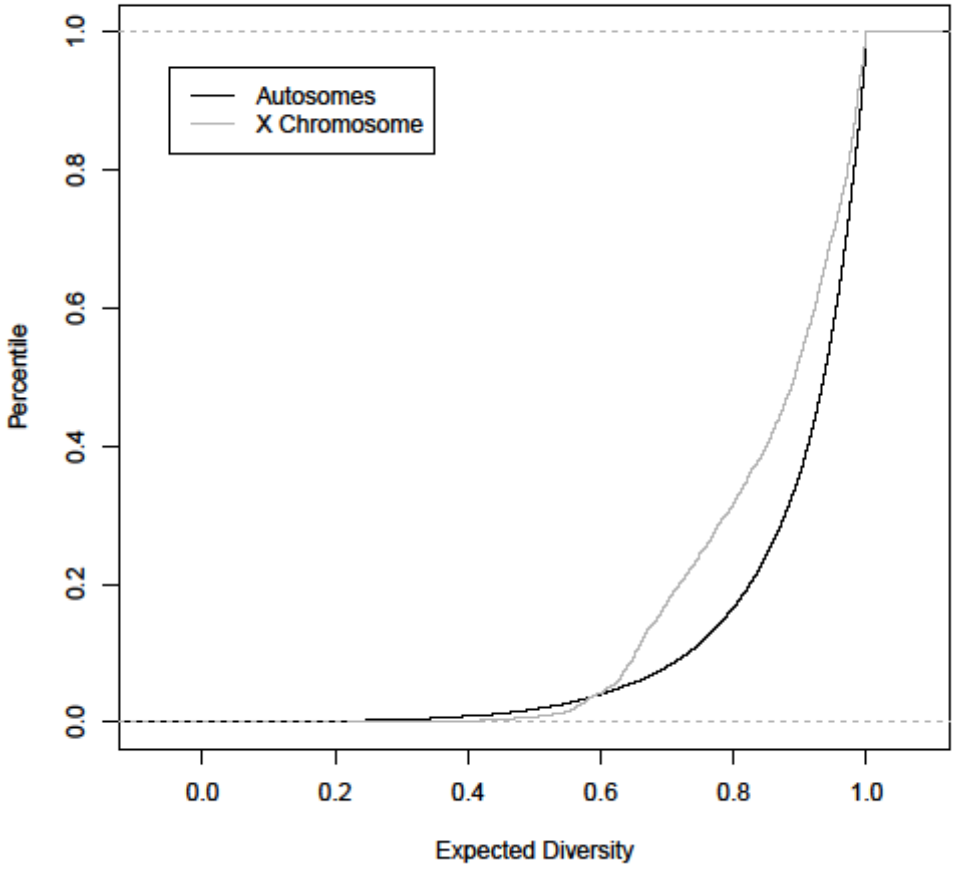


Figure 5. Scatterplot of YRI π/D versus the local recombination rate. The gray dots represent loci in the neutralome in the autosomes (left) and X chromosome (right). Each line represent the IRLS regression line with 95% CI fit to the neutralome (solid) and the nongenic portion of the genome (dashed).

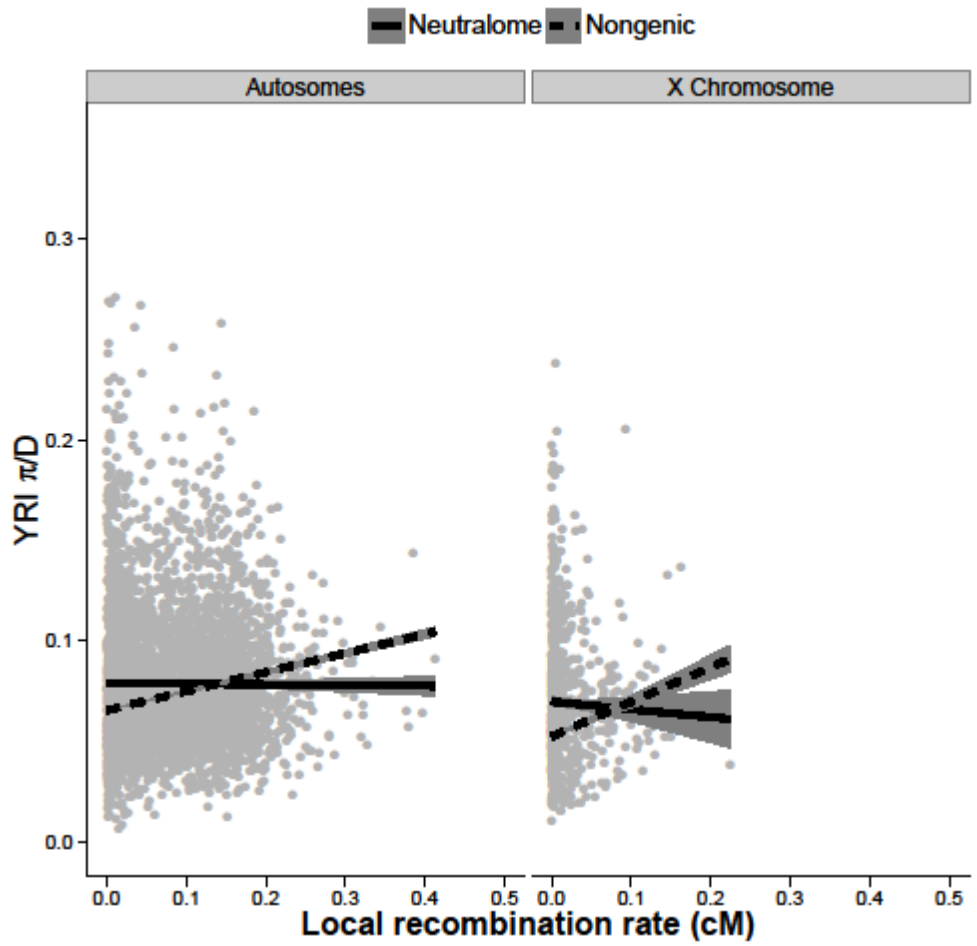


Figure 6. Mean π/D (± SEM) in several regions of the autosomes (left) and the X chromosome (right). NRE loci are regions predicted to be neutral by Neutral Region Explorer, while the neutralome refers to loci identified in this study. - refers to the set difference, ∩ is the intersection, and the neutralome and the NRE loci have some overlap. Statistical comparisons between NRE and the neutralome were made between nonoverlapping sets (Neutralome - NRE versus NRE – Neutralome). Genic and nongenic loci are also included for comparison purposes. Lower π/D is consistent with higher levels of background selection and/or genetic hitchhiking.

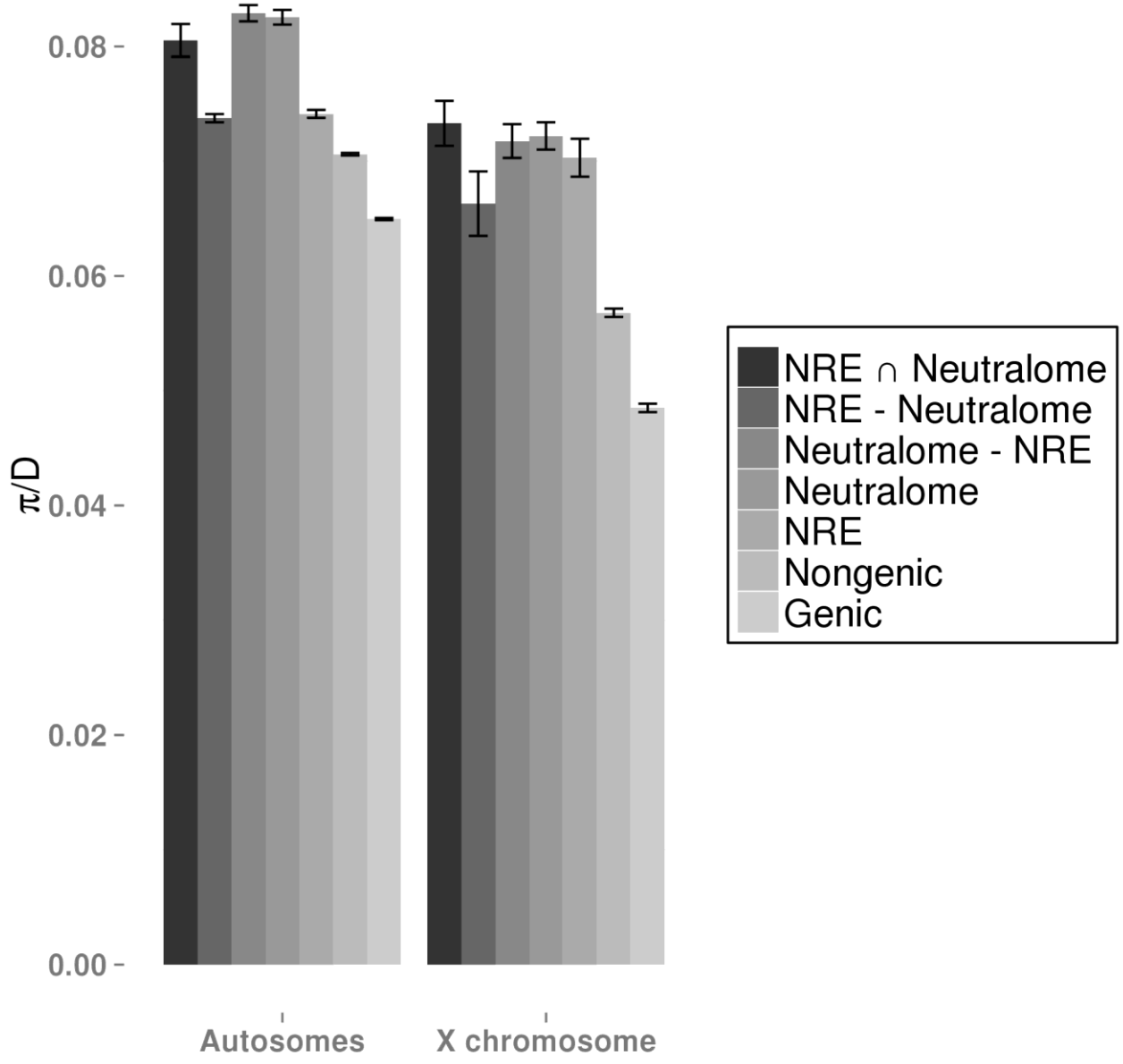
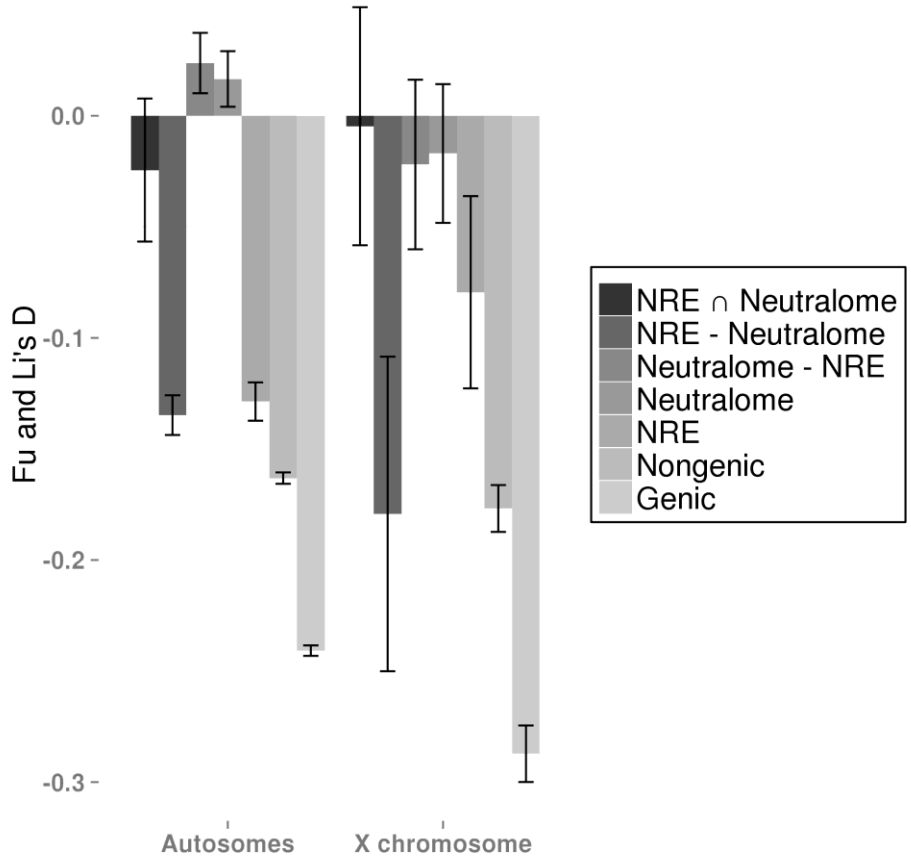


Figure 7. Mean Fu and Li's D (± SEM) in several regions of the autosomes (left) and the X chromosome (right). NRE loci are regions predicted to be neutral by Neutral Region Explorer, while the neutralome refers to loci identified in this study. - refers to the set difference, ∩ is the intersection, and the neutralome and the NRE loci have some overlap. Statistical comparisons between NRE and the neutralome were made between nonoverlapping sets (Neutralome - NRE versus NRE – Neutralome). Genic and nongenic loci are also included for comparison purposes. Relative to a neutral model with no growth, negative Fu and Li's D signifies an excess of singletons, while positive values indicate a reduction in singletons. Weak background selection and genetic hitchhiking may both lead to negative Fu and Li's D, as will demographic effects such as population growth.



TABLES

Table 1. Beta coefficients and confidence intervals inferred from modeling $\pi/D \sim R + G$ on the X chromosome (X) and the autosomes (A). NeX/NeA refers to the relative effective population size of the X chromosome versus the autosomes.

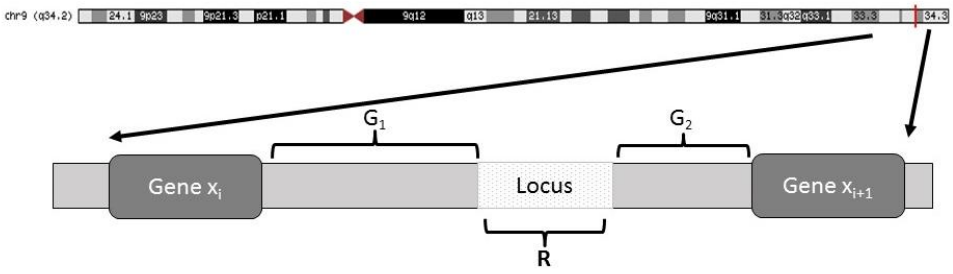
NeX/NeA	Population	Parameter	β coefficient from IRLS regression	Lower CI (0.025)	Upper CI (0.975)	*p*-value between rows
0.75	YRI	X_G	0.32	0.28	0.36	< 0.001
0.75	YRI	X_R	0.15	0.12	0.19	< 0.001
0.75	YRI	A_R	0.07	0.07	0.08	< 0.001
0.75	YRI	A_G	0.04	0.03	0.04	
0.75	CEU	X_G	0.20	0.16	0.24	< 0.001
0.75	CEU	X_R	0.10	0.07	0.13	0.002
0.75	CEU	A_R	0.05	0.05	0.06	< 0.001
0.75	CEU	A_G	0.03	0.03	0.03	
0.95	YRI	X_G	0.26	0.23	0.29	< 0.001
0.95	YRI	X_R	0.12	0.10	0.15	< 0.001
0.95	YRI	A_R	0.07	0.07	0.08	< 0.001
0.95	YRI	A_G	0.04	0.04	0.04	
0.95	CEU	X_G	0.16	0.13	0.19	0.001
0.95	CEU	X_R	0.08	0.05	0.10	0.028
0.95	CEU	A_R	0.05	0.05	0.06	< 0.001
0.95	CEU	A_G	0.03	0.03	0.04	

Table 2. Demographic parameters estimated considering three definitions of neutral sites. Sites from the neutralome, sites identified by neutral region explorer (NRE), as well as 4-fold degenerate sites were used infer 2-epoch instantaneous population growth model in δaδI. δaδI was then used to estimate 3 parameters: the ancestral population size (Na), the rate of growth (the contemporary effective population size/Na), and the time of growth (*T*; given in thousands of years). Confidence intervals, as well significant differences in the parameters between these neutral definitions, were ascertained using a 1000-iteration bootstrap.

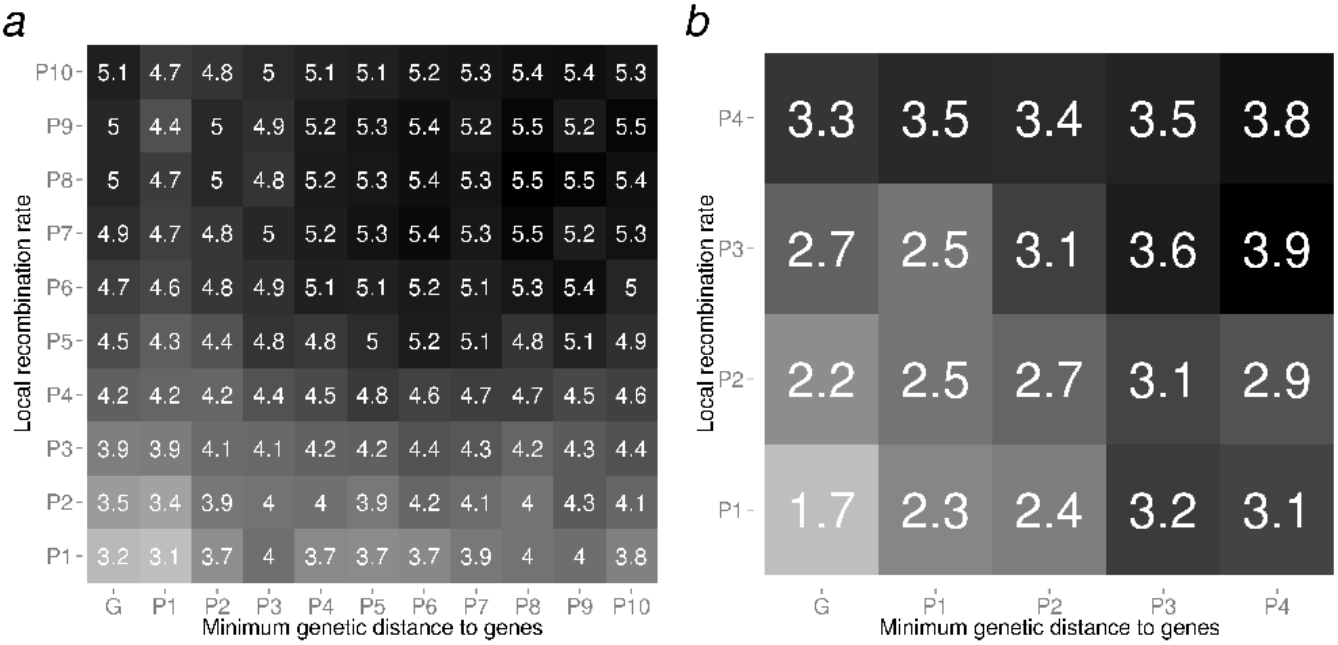
Sample	Parameter	Point Estimate	0.025 CI	0.975 CI	*p*-value (between rows)
Neutralome	Na	11,332	10,522	11,746	0.001
NRE	Na	8,918	8,716	9,093	< 0.001
4-Fold	Na	5,657	4,941	6,099	
Neutralome	*n*	1.59	1.55	1.68	< 0.001
NRE	*n*	1.82	1.80	1.85	< 0.001
4-Fold	*n*	2.57	2.42	2.84	
Neutralome	T (kya)	161.76	128.10	220.66	0.024
NRE	T (kya)	125.53	113.98	138.49	0.040
4-Fold	T (kya)	94.25	73.23	127.40	

SUPPLEMENTAL FIGURES

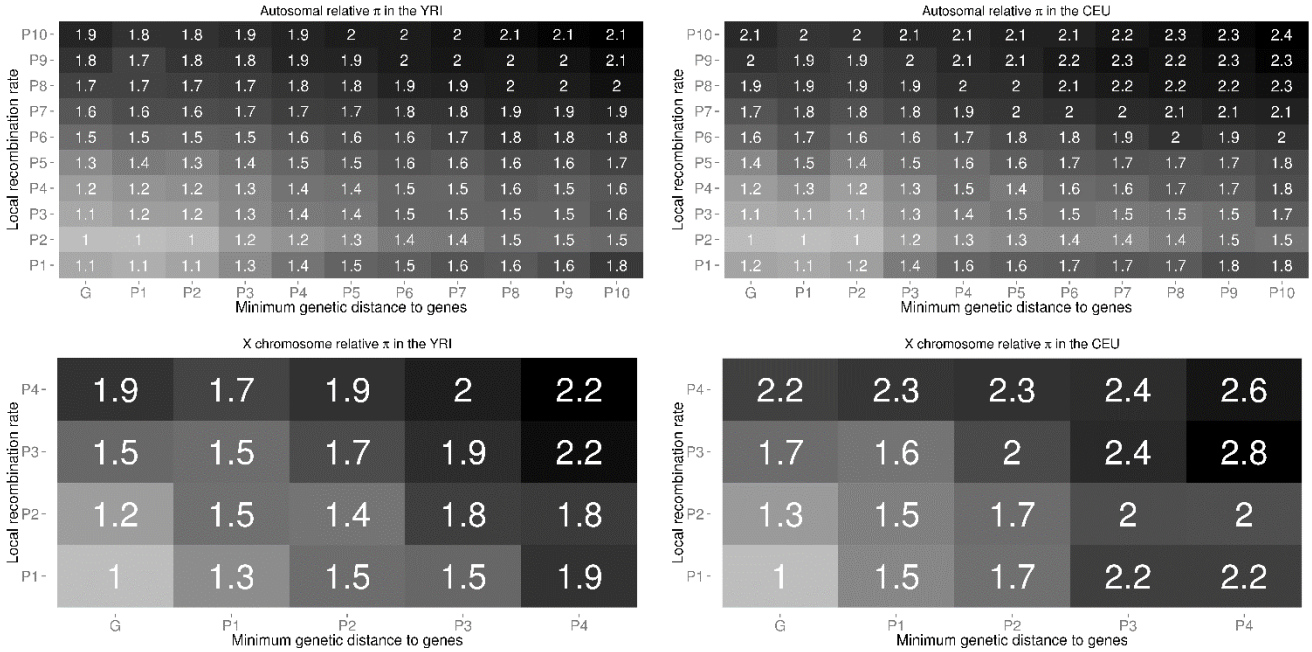
Supplemental Figure 1. Measurements on a single locus used in this study. We partitioned each chromosome (karyotype on top) into 10kb loci (blow-up, bottom). Within each locus we computed both diversity (π) within populations and divergence (D) between each population and the inferred ancestor of orangutans. We also computed the rate of recombination (R) within each locus, as well as the genetic distance to genes, G_1 and G_2, in centimorgans (cM). We took the minimum genetic distance (G) as the minimum of G_1 and G_2. Note that $R << G$.



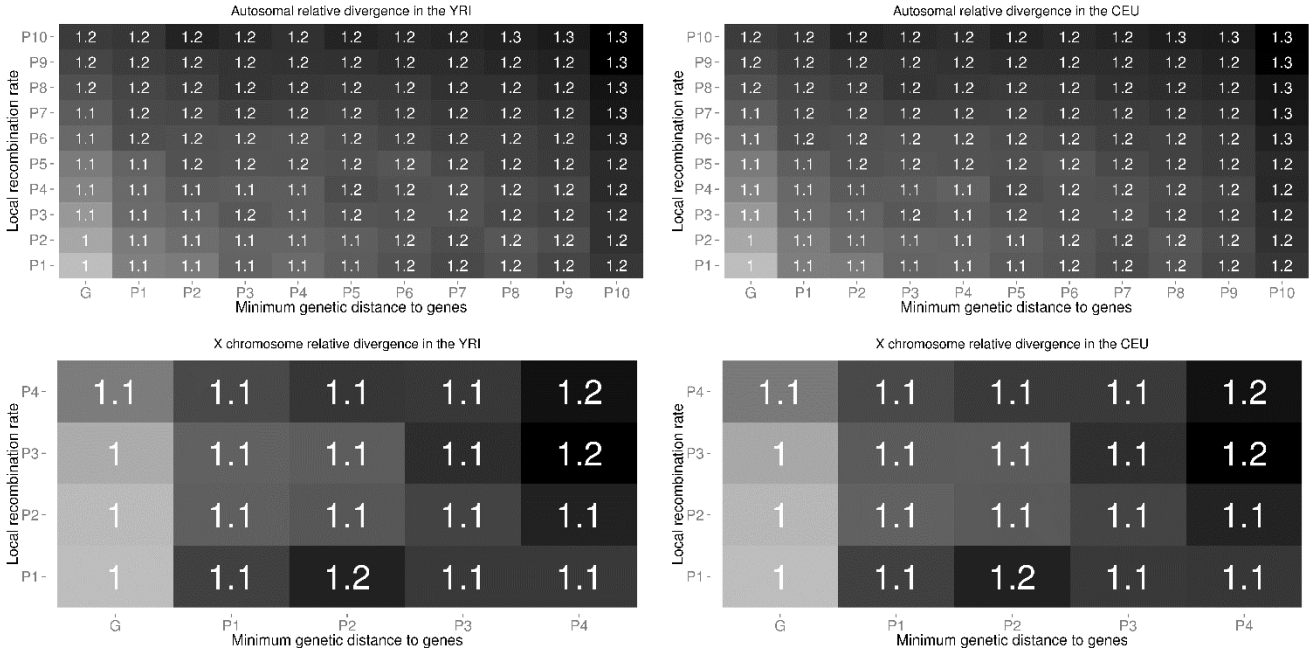
Supplemental Figure 2. Heatmap of the median π/D (x 100) in CEU. The x-axis shows the minimum genetic distance to genes and the y-axis the local recombination rate. Darker cells correspond to higher π/D. Each cell corresponds to a pair of percentiles (P) in the distance to genes and the local recombination rate, with column G corresponding to loci in genes, and the remaining columns being outside of genes. a) Autosomes. b) X chromosome



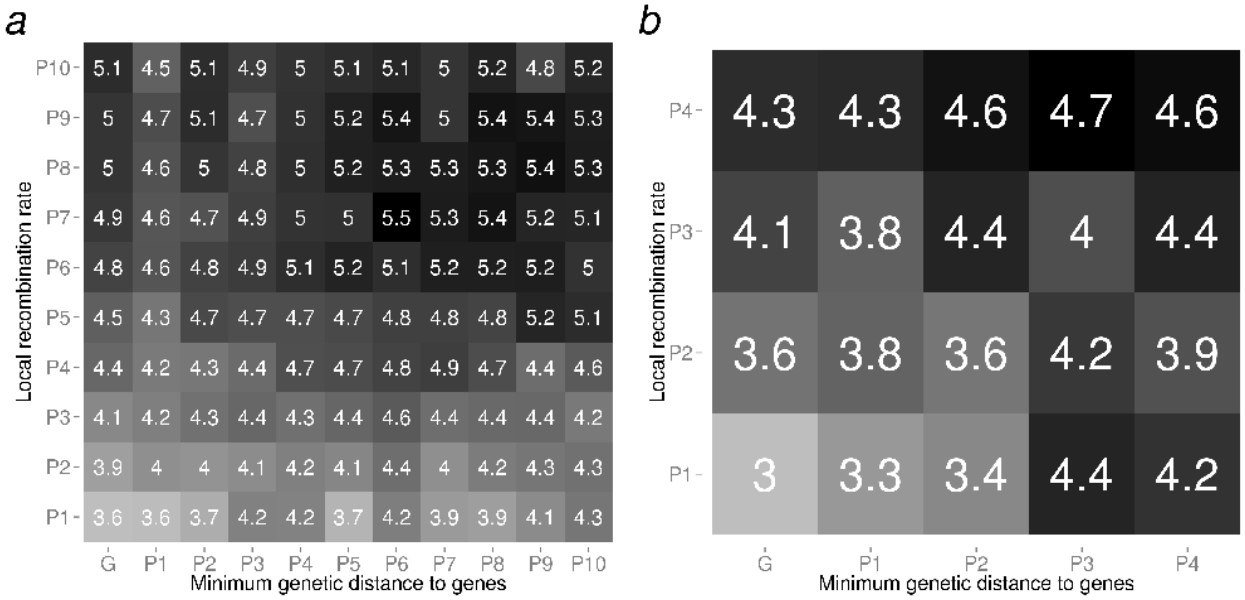
Supplemental Figure 3. Heatmap of the median relative π (divided by the minimum bin) versus the minimum genetic distance to genes (x-axis, binned by percentiles) and the local recombination rate (y-axis, binned by percentiles), in the YRI (left) and CEU (right) for the autosomes (top) and X chromosome (bottom). Darker cells correspond to higher relative π. Each cell corresponds to a pair of percentiles in the distance to genes and the local recombination rate, with the value shown being the median for loci in that bin.



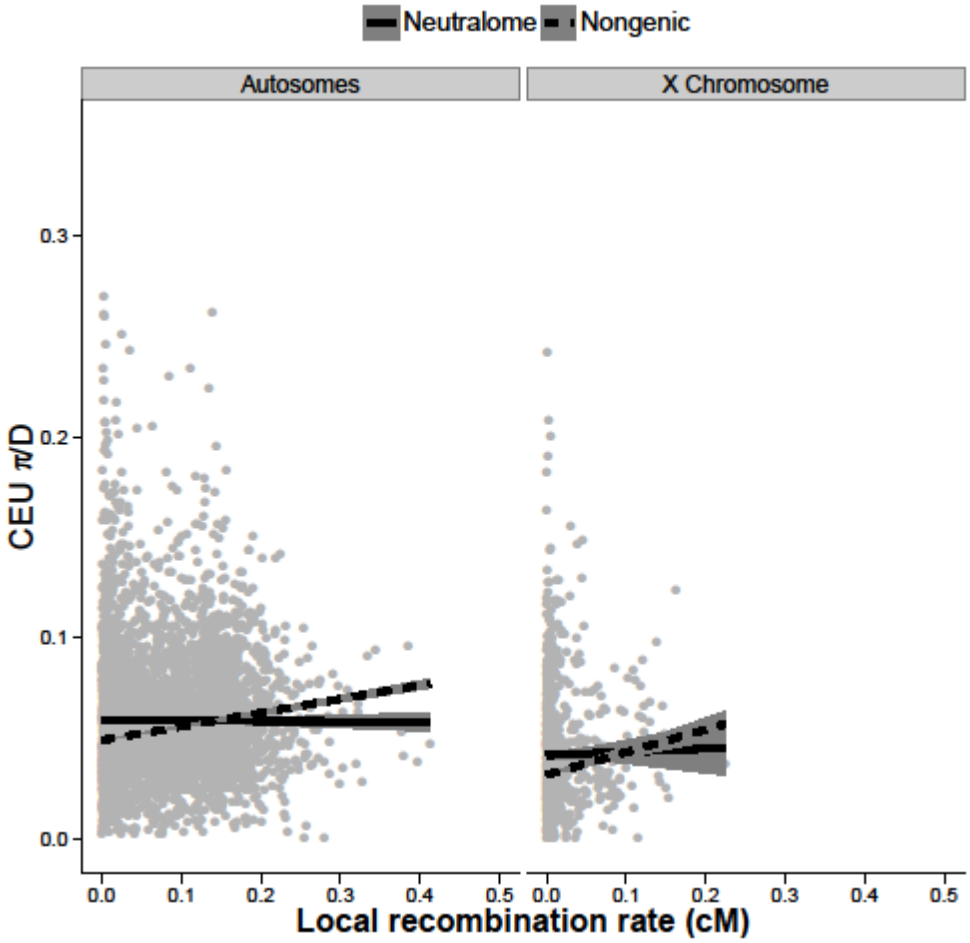
Supplemental Figure 4. Heatmap of the median relative divergence (divided by the minimum bin) versus the minimum genetic distance to genes (x-axis, binned by percentiles) and the local recombination rate (y-axis, binned by percentiles), in the YRI (left) and CEU (right) for the autosomes (top) and X chromosome (bottom). Darker cells correspond to higher divergence. Each cell corresponds to a pair of percentiles in the distance to genes and the local recombination rate, with the value shown being the median for loci in that bin.



Supplemental Figure 5. Heatmap of the median π/D (x 100) in CEU considering only W→W and S→S transversions. The x-axis shows the minimum genetic distance to genes and the y-axis the local recombination rate. Darker cells corresponding to higher π/D. Each cell corresponds to a pair of percentiles (P) in the distance to genes and the local recombination rate, with column G corresponding to loci in genes, and the remaining columns being outside of genes. A) Autosomes. B) X chromosome



Supplemental Figure 6. Scatterplot of CEU π/D versus the local recombination rate. The gray dots represent loci in the neutralome in the autosomes (left) and X chromosome (right). Each line represent the IRLS regression line with 95% CI fit to the neutralome (solid) and the nongenic portion of the genome (dashed).



Supplementary Table 1. Beta coefficients and confidence intervals inferred from modeling π/D ~ R + G on the X chromosome (X) and the autosomes (A) considering only W->W and S->S transversions. NeX/NeA refers to the relative effective population size of the X chromosome versus the autosomes.

NeX/NeA	Population	Parameter	β coefficient from IRLS regression	Lower CI (0.025)	Upper CI (0.975)	*p*-value between rows
0.75	YRI	X_G	0.11	0.08	0.13	0.059
0.75	YRI	X_R	0.08	0.05	0.1	< 0.001
0.75	YRI	A_R	0.02	0.02	0.03	< 0.001
0.75	YRI	A_G	0.01	0.01	0.02	
0.75	CEU	X_G	0.03	0.01	0.06	0.76
0.75	CEU	X_R	0.05	0.02	0.07	0.003
0.75	CEU	A_R	0.02	0.01	0.02	< 0.001
0.75	CEU	A_G	0.01	0	0.01	
0.95	YRI	X_G	0.08	0.06	0.11	0.06
0.95	YRI	X_R	0.06	0.04	0.08	< 0.001
0.95	YRI	A_R	0.02	0.02	0.03	< 0.001
0.95	YRI	A_G	0.01	0.01	0.02	
0.95	CEU	X_G	0.04	0.01	0.05	0.742
0.95	CEU	X_R	0.03	0.02	0.06	0.018
0.95	CEU	A_R	0.02	0.01	0.02	< 0.001
0.95	CEU	A_G	0.01	0.01	0.01	

Supplementary Table 2. The top 100 best fittng parameter values for Model C of Halligan et al. (2013). The sum of squared errors (SS), along with the 5 parameters (P1-P5) used to fit the observed levels of diversity to the distribution of phastCons elements are shown for the autosomes (left) and the X chromosome (right).
Too large for reproduction. See online material

Supplementary Table 3. Contrasting the mean of several summary statistics in the neutralome vs. regions found from Neutral Region Explorer (NRE). The means for different summary statistics (rows) were computed for regions found by NRE and not the neutralome (NRE-Neutralome) and for regions found by the Neutralome and not NRE (Neutralome - NRE). Welch t-tests were used to assess the significance of the difference in means across summary statistics.

Summary Statistic	Compartment	Mean(NRE – Neutralome)	Mean(Neutralome – NRE)	P-value (1-tailed)
Θ_π/D	Autosomes	0.074	0.083	< 2.2e-16
Θ_π/D	X chromosome	0.066	0.072	0.043
Θ_W/D	Autosomes	0.079	0.087	< 2.2e-16
Θ_W/D	X chromosome	0.071	0.075	0.080
Tajima's D	Autosomes	-0.333	-0.257	2.148E-08
Tajima's D	X chromosome	-0.241	0.019	0.046
Fu & Li's D	Autosomes	-0.135	0.024	< 2.2e-16
Fu & Li's D	X chromosome	-0.179	-0.022	0.026
TMRCA	Autosomes	0.156	0.180	< 2.2e-16
TMRCA	X chromosome	0.134	0.150	0.014

Supplemental Table 4. Contrasting the mean of several summary statistics in the neutralome vs. regions found from Neutral Region Explorer (NRE). The means for different summary statistics (rows) were computed for regions found by NRE and not the neutralome (NRE-Neutralome) and for regions found by the Neutralome and not NRE (Neutralome - NRE). To control for NRE generating more "neutral" loci, the background coefficient of McVicker et al. (2009) (BG) was varied from 0.95 to 0.98, which gave roughly comparable sample sizes between the NRE loci and the Neutralome. Note that in doing so the sample size on the X chromosome for the NRE loci went from 170 to 21, making statistical power on the X chromosome problematic for these comparisons. Welch t-tests were used to assess the significance of the difference in means across summary statistics.

Summary Statistic	Compartment	Mean(NRE – Neutralome)	Mean(Neutralome – NRE)	P-value (1-tailed)	
Θ_π/D	Autosomes	0.077	0.083	6.675E-12	
Θ_π/D	X chromosome	0.060	0.072		0.069
Θ_W/D	Autosomes	0.081	0.087	3.888E-15	
Θ_W/D	X chromosome	0.067	0.076		0.095
Tajima's D	Autosomes	-0.271	-0.257		0.203
Tajima's D	X chromosome	-0.465	-0.209		0.046
Fu & Li's D	Autosomes	-0.047	0.025	1.013E-04	
Fu & Li's D	X chromosome	-0.223	-0.023		0.186
TMRCA	Autosomes	0.164	0.179	5.075E-14	
TMRCA	X chromosome	0.110	0.152		0.012

APPENDIX B

GENOMIC INFERENCE ON SEXUAL SELECTION IN THE GREAT APES

Manuscript prepared for submission to the Proceedings of the National Academy of Sciences

Intended as a research article at the Proceedings of the National Academy of Sciences

Genomic Inference on Sexual Selection in the Great Apes

Authors:
August E Woerner[1,2], Krishna R Veeramah[3], Joseph C Watkins[4], Laurie S Stevison[5], Jeffrey D Wall,[6], Michael F Hammer[1*]

Affiliations:

[1]Arizona Research Laboratories Division of Biotechnology, [2]Department of Genetics, [4]Department of Mathematics, University of Arizona, Tucson AZ 85721.

[3]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794.

[5]Department of Biological Sciences, Auburn University, Auburn AL. 36849.

[6]Department of Human Genetics, University of California San Francisco, San Francisco CA

**Correspondence should be addressed to:*

Michael F. Hammer, PhD

ARL Division of Biotechnology, 111K Keating Building, 1657 E. Helen Street, University of Arizona, Tucson, AZ, 85721, USA

Phone: +1-520- 621-9828

e-mail: mfh@email.arizona.edu

Keywords: Linked selection, neutralome, sex ratio, great apes, sperm competition

ABSTRACT

Great ape species exhibit a range of mating strategies associated with different levels of sexual selection, ranging from multimale-multifemale mating in bonobos and chimpanzees, to mild polygyny in humans, to a unimale mating system in gorillas. These systems predict different relative levels of nucleotide polymorphism on the X chromosome and autosomes. These chromosomes are also influenced by variable rates of natural selection acting on both coding and noncoding sites. Here we analyze whole genome sequence data from 9 humans, 13 bonobos, 10 Nigerian chimpanzees, and 14 western gorillas to evaluate the relative impact of sexual and natural selection on diversity and patterns of linkage disequilibrium (LD) on the autosomes and X chromosome. We show genome-wide reductions in diversity (likely resulting from purifying and positive selection), the extent of which varies considerably by taxon and chromosome. Interestingly, <1% of the genomes of all four species are predicted to be unaffected by selection. Using these sites in each species we infer the effective population size of females (N_f) and males (N_m) based on patterns of nucleotide variation. We applied a second approach to infer N_f/N_m using genetic maps. We find a statistically significantly skewed sex ratio in all four species, with values of N_f/N_m of ~2:1 in humans, bonobos and gorillas, and a breeding sex ratio as high as 10:1 in chimpanzees. These estimates suggest that unimale mating may not be strict in gorillas, and that bonobos and chimpanzees differ in rates of sperm competition and/or behaviors that lead to higher variances in male reproductive success.

INTRODUCTION

Both sexual selection and natural selection play a considerable role in shaping genomic patterns of diversity on the X chromosome and the autosomes. The X chromosome spends 2/3 of its time in females (but see Goldberg and Rosenberg 2015), and thus it is differentially impacted by the sexes. In the absence of natural selection, the effective population size of the X chromosome (Ne_X) relative to the autosomes (Ne_A) is an estimator of the number of breeding females (N_f) relative to males (N_m) in a population. Under standard assumptions of neutrality and a sex-ratio of 1:1, the expected Ne_X/Ne_A is 0.75 in simple models of random mating (Hedrick 2007). Under more complicated scenarios this ratio can vary tremendously (Caballero 1995), reaching a maximum of 9/8 (Hedrick 2007), assuming discrete generations. Demographic forces such as population expansions and contractions may also influence Ne_X/Ne_A (Pool and Nielsen 2007), though their effects are ephemeral. Molecular estimates of Ne_X/Ne_A in great apes are of particular importance as they exhibit a diverse range of mating systems that generally predict $N_f/N_m > 1$. Gorillas have extreme sexual dimorphism, and have a polygynous mating system with a single dominate male (Dixson 1998). Chimpanzees and bonobos have moderate sexual dimorphism and have a multi-male, multi-female system (Dixson 1998) wherein sexual selection is thought to operate at the level of the genitalia through sperm competition (Harcourt et al. 1981), though behavioral variations between chimpanzees and bonobos might also contribute to N_f/N_m (Furuichi 2011) (Table 1). And while more contentious, human populations are generally regarded as moderately polygynous (Low 1988). Despite these expectations, increased levels of natural selection on the X chromosome relative to the autosomes biases estimates of N_f/N_m downwards (e.g. < 1) in great apes (e.g. Keinan et al. 2009; Hammer et al. 2010; Hvilsom et al. 2012; but see Prado-Martinez et al. 2013).

While the X chromosome and the autosomes have differences in gene expression (Lercher et al. 2003; Nguyen et al. 2015), gene content (Wang et al. 2001), and in recombination rate (Frazer et al 2007; Roach et al. 2010), the primary differences between these systems is hemizygosity (Hvilsom et al. 2012;

Veeramah et al. 2014, but see Nguyen et al. 2015). Hemizygosity in males exposes recessive alleles on the X chromosome to selection. Thus recessive allele deleterious alleles are removed more efficiently and recessive advantageous alleles are more likely driven to fixation (Charlesworth et al. 1987). Consistent with these predictions, rates of adaptive amino-acid substitution are higher on the X chromosome than on the autosomes in humans (Nielsen et al. 2005; Veeramah et al. 2014) and in central chimpanzees (Hvilsom et al. 2012).

Nearby sites on the chromosome are often linked, and as such neutral alleles can be driven to fixation with genetic hitchhiking (Maynard Smith and Haigh 1974) or extinction through background selection (Charlesworth et al. 1993) when linked to positively or negatively selected alleles, respectively. Thus, when compared to the autosomes the effects of genetic hitchhiking are predicted to be greater on the X chromosome (Maynard Smith and Haigh 1974; Begun and Whitley 2000), while the effects of background selection are predicted to be lesser (Charlesworth 1996). Within the context of great apes these predictions are consistent with greater reductions in diversity on the X chromosome than on the autosomes (Hammer et al. 2010; Gottipati et al. 2011; Prado-Martinez et al. 2013; Arbiza et al. 2014) due to increased rates of hitchhiking (Nam et al. 2015). Differing levels of linked genic selection between the X chromosome and the autosomes biases inferences on N_f/N_m (i.e. Ne_X/Ne_A) in great apes that use the entirety of these genomic compartments (Keinan et al. 2009; Hammer et al. 2010; Prado-Martinez et al. 2013).

Genes are neither the sole targets of selection, nor are they most strongly selected elements in the genome (Bejarno et al. 2004; Halligan et al. 2011). Phylogenetically inferred conserved elements (CEs) are small (~30-200bp) loci that show far fewer substitutions than would be expected by neutral sequence. CEs may number in the millions and are more often noncoding than coding (Bejarno et al. 2004; Cooper et al. 2005; Siepel et al. 2005). Different classes of the CEs appear to be either under negative selection (Katzman et al. 2007; Chen et al. 2007; McVicker et al. 2009) or under mixtures of positive and negative

selection (Torgerson et al. 2009; Halligan et al. 2011). Conserved elements reduce diversity at linked sites (Hernandez et al. 2011; Halligan et al. 2013), profoundly reducing estimates of effective population size across the human genome (Woerner et al. 2016).

Diversity-based estimates of Ne_X/Ne_A are biased downwards in great apes when we consider either the whole of the X chromosome versus the whole of the autosomes or when we consider regions near genes (Hammer et al. 2010; Prado-Martinez et al. 2013). As the X chromosome has a lower gene-density than the autosomes (Deloukas et al. 1998), and as some CEs are regulatory (Pennacchio et al. 2006), the X chromosome may have a lower density of CEs than the autosomes (but see Davydov et al. 2010). Thus, linkage to nongenic CEs may also bias Ne_X/Ne_A even in regions far from genes (Woerner et al. 2016).

To this end, this paper estimates N_f/N_m in a broad sampling of great apes. We estimate N_f/N_m using whole genome sequence data from 13 bonobos, 10 Nigerian chimpanzees, and 14 western gorillas (Prado-Martinez et al. 2013) and 9 sub-Saharan humans (the Yoruba, of Ibadan Nigeria) (Drmanac et al. 2010), in conjunction with recently developed genetic maps in non-human great apes (Stevison et al. 2016). We parse the effects of linked selection on diversity into genic and nongenic sources by measuring both levels of linkage to the nearest gene and the local rates of recombination (as per Woerner et al. 2016). As phylogenetically conserved *phastCons* elements play a substantial role in shaping patterns of diversity in humans (Woerner et al. 2016), we explore how this conserved feature of the genome varies between the X chromosome and the autosomes. We then model the genomic effects of linkage to various classes of *phastCons* elements and contrast levels of linked selection across the autosomes and the X chromosome, as well as across taxa. We then evaluate whether these linked effects vary between the X chromosome and the autosomes, and if they depend on linkage to genes. Next, we extend the work of Woerner et al. (2016), and introduce a general framework for inferring the locations of the genome least-affected by selection (i.e. neutralomes) based on genomic modeling of linked selection to *phastCons*

elements and information on genes. We use neutralomes to infer Ne_X/Ne_A across our taxa, and in comparing them to other classes of "neutral" we assess how assumptions on (a lack of) selection can fundamentally alter genomic inferences. We then compare these neutralome's Ne_X/Ne_A to estimates from genetic maps, and compare how N_f/N_m contrasts to contemporary mating systems in great apes.

RESULTS

Diversity, linkage to genes, and the local recombination rate.

We estimated diversity and linkage at fine-scales using resequencing data in four great ape species. All computations involved two sets of loci; nonoverlapping 10kb loci that span the genome used in our exploratory analyses (herein termed 10kb loci), and nonoverlapping nongenic, non-*phastCons* element loci with exactly 2kb of called sequence that span no more than 10kb, used in our inferential analyses (herein termed 2kb loci). After masking out repeats and uncallable sequence we computed nucleotide diversity (π) and divergence (D) to an ancestral sequence. Next we assessed levels of linkage, computing the recombination rate R in centimorgans (cM) within each locus, as well as the minimum distance to genes G in cM.

We visualized the joint effects of R and G on diversity using deciles in the autosomes (top) and quartiles for the X chromosome (bottom) (Fig. 1) for each ape species (left to right) in our 10kb loci. With D controlling for heterogeneity in mutation rate, we display the median π/D for each bin. Consistent with previous works, diversity increases with the genetic distance to genes (within each plot; left to right columns) in humans (Hammer et al. 2010; Gottipati et al. 2011; Arbiza et al. 2014) and across great apes (Prado-Martinez et al. 2013). π/D also increases with R regardless of G (bottom to top rows), with linkage to phylogenetically conserved *phastCons* elements driving this trend in humans (Woerner et al. 2016). The relationships among π/D, R and G differ across species, with bonobos showing little effects for G, while gorillas show extremely stark gradients in G. Humans and chimpanzees display similar patterning

on both the X chromosome and the autosomes. The lowest levels of diversity are often not in genes (column G), but in low-recombination gene-adjacent regions (column P1, rows P1 and P2), which may indicate pronounced selection at regulatory sites.

To assess the statistical significance of the observed trends we used iterated reweighted least squares (IRLS) regression to model $\pi/D \sim R + G$ for the autosomes and the X chromosome separately across taxa using our 2kb loci. IRLS regression estimates standardized slope coefficients (β), which are measures of effect size. The estimated β coefficients are significantly greater than 0 across taxa on both the autosomes and the X chromosome (Table S1, $p<0.001$ in all comparisons except for β for linkage to genes on the X chromosome in bonobo, $p = 0.049$). Across all taxa save bonobo, β-coefficients are higher on the X chromosome than the autosomes (Fig. 2, Table S1), with the β for linkage to genes on the X chromosome (β_{XG}) being the overall highest. Bonobos, on the other hand, have only small slopes for linkage to genes on both the autosomes and the X chromosome (Fig. 2). The order of autosomal coefficients (β_A) vary between gorillas and the rest of the taxa, with $\beta_{AG} < \beta_{AR}$ ($p < 0.001$ across all comparisons) in humans, chimps and bonobos, but $\beta_{AR} < \beta_{AG}$ ($p < 0.001$) in gorillas. Compared to other taxa, β_{AG} appears elevated in gorillas, while β_{AR} appears relatively constant.

The density of genes and conserved elements on the X chromosome versus the autosomes

The correlation between R and π/D in humans is driven by linked selection to phylogenetically conserved *phastCons* elements (Woerner et al. 2016). *phastCons* elements are short (~30-200bp), interspersed elements that are common outside of genes and have both sequence constraint (Siepal et al. 2005; Hernandez et al. 2011; Halligan et al. 2011, 2013) and regulatory function (Pennacchio et al. 2006). We investigated the densities of *phastCons* elements (SI Materials and Methods) and genes on the X chromosome versus the autosomes to better understand their selective roles in these two systems. Since the X chromosome has less genes than the autosomes (Deloukas et al. 1998), it may also have less

phastCons elements. Recapitulating Deloukas et al. (1998) but with current annotations (human genes from Ensembl build 72), we find that the X chromosome has significantly lower gene-density (34% of bases) than the autosomes (47% of bases) ($p < 2.2e\text{-}16$, χ^2 test). We considered three classes of *phastCons* elements: Elements inferred in placental mammals (mammals), elements inferred in primates (primates), and base-positions that were considered conserved in primates that were naively clustered (primate clusters). Primate elements, mammal elements, and primate clusters each have lower densities on the X chromosome (2.6, 2.8 and 2.6%, respectively) than the autosomes (3.7, 4.1 and 3.7%, respectively) ($p < 2.2e\text{-}16$ across all comparisons, χ^2 test).

Genomic estimates of the local effective population size

To better quantify reductions in diversity due to linkage to *phastCons* elements, we categorized our *phastCons* elements into those that intersect exons (CEEs) and those that do not (CNEs), and modeled their genomic effects. Using Model C of Halligan et al. (2013) we estimated the cumulative effects of linkage to both CNE and CEE elements (SI Materials and Methods), forming estimates of the relative effective population size (ν) for each of our 2kb loci. ν is a prediction on π/D that measures the combined reductions in diversity that stem from linkage to all *phastCons* elements within a 1 cM window around each locus, with this estimate given relative to a neutral rate of 1.0 (Woerner et al. 2016). We fit π/D separately for the autosomes and the X chromosome, and separately considering our three classes of *phastCons* elements in each taxon under a sum of squares criterion. Assessments of model fit, measured as the coefficient of determination (R^2), indicate better fit for the X chromosome than the autosomes across all taxa (Table S2). Bonobos have the worst model fits, while gorillas have the best model fits. In general, primate *phastCons* elements yielded the lowest R^2, mammalian *phastCons* elements yielded intermediate R^2, and primate *phastCons* clusters yielded the highest R^2. Given this we selected the ν inferred using primate clusters as our final estimator of ν unless stated otherwise.

Evaluation of the cumulative distribution of ν (Fig. 3) shows that despite the fact that the X chromosome has lower gene/*phastCons* densities, it has a larger proportion of nucleotides affected by linked selection for all taxa save bonobos. To assess if the differences in ν between the X chromosome and the autosomes are due to differences in local linkage disequilibrium (LD), we selected matching loci (on *R*) between the X chromosome and the autosomes, creating a paired sample with approximately the same local rate of recombination. Both matching on *R* (Table S3*A*), and using the entirety of the X chromosome and the autosomes (Table S3*B*), yielded a mean ν higher on the autosomes than the X chromosome for gorillas, Nigerian chimpanzees and the Yoruba. The opposite is true for bonobos ($p < 2.2e-16$ across all comparisons, paired and unpaired 2-sample *t*-test). Repeating both procedures in regions far from genes (>0.3 cM) we see a substantively different picture. When constrained to be far from genes, the X chromosome instead shows *less* constraint (i.e. higher ν) than the autosomes in all taxa both when we match on *R* ($p < 1e-15$ across all comparisons) and when we do not ($p < 2.2e-16$ across all comparisons) (Table S4).

To visualize the transition-point on the X chromosome versus the autosomes we plotted ν as a function of *G* across taxa (Fig. 4), smoothing ν with a generalized additive model. In all taxa ν is lower near genes on the X chromosome than on the autosomes, consistent with greater selection near genes (Hammer et al. 2010; Prado-Martinez et al. 2013). However, in each taxon the value for *G* where the curves in Fig. 4 cross indicates a transition from more to less constraint on the X chromosome relative to the autosomes. Further, the transition points vary between taxa, with the change in bonobos occurring immediately after genes (~0 cM), humans and chimpanzees transitioning at medial distance from genes (~0.15 cM), and the transition in gorillas occurring last (~0.3 cM). π/D was estimated in the same manner as ν to assess if the predictions of diversity (ν) correspond to diversity itself (Fig. S1). Though the interpretation is more difficult, π/D appears to be inflated in regions especially far (i.e., essentially unlinked) from genes.

Inference of neutralomes

To form more correct estimates of the neutral Ne_X/Ne_A we developed an extension of the heuristic of Woerner et al. (2016) to find "neutralomes", i.e., loci in which ν, *R* and *G* are no longer significantly correlated with π/D (SI Materials and Methods). Like in Woerner et al. (2016), our heuristic finds breakpoints in ν, and with our extension it also finds breakpoints in ν and *G*, to form candidate sets of neutral regions. Finding proper thresholds can be difficult, particularly given that assessments of neutrality hinge upon statistical power which is a function of sample size. We address this by finding a threshold using a constrained piecewise (segmented) regression. Using an iterative combination of linear and nonlinear regression techniques (SI Materials and Methods) we inferred neutralomes solely on the basis of ν (Table S5*A*), and by thresholding on both *G* and ν (Table S5*B*). As conditioning on both *G* and ν generally led to the inference of larger neutralomes we consider the loci in Table 3*B* as the neutralomes in each taxa, though we note that the reliance on *G* requires additional information whose quality is likely to improve with higher quality genome sequencing. Methods that threshold solely on ν fail to infer the presence of a neutralome of sufficient size (we consider a minimum sample size of 50) in the autosomes of gorillas and bonobos. When also conditioned on *G*, not only does the size of almost every neutralome increase (excepting the X chromosome of gorillas), but we also find neutralomes in taxa where the procedure failed when solely considering ν (Table S5*A*).

Validation of neutralomes

The inference of neutralomes is impacted by a lack of power to detect relationships between π/D, ν, *G* and *R*. To address this concern we combined neutralomes across the autosomes and the X chromosome in all taxa into a single dataset of 20,788 loci. After normalizing terms we used IRLS regression to model $\pi/D \sim \nu + G + R$ in our combined dataset, which yielded coefficient estimates $\beta_\nu = 0.023$ (p=0.69), β_G =-0.012

(p=0.03) and a β_R = -0.0004 (p=0.94), with an overall regression F-statistic of 1.543 (p=0.20). Note that the sign of β_G (and β_R) is opposite of our *a priori* expectations, which inflates the F-statistic. To assess for significant monotonic relationships we used 1-tailed Kendall's rank-correlations to test for significant positive correlations between π/D and ν (τ = 0.002, p= 0.32), π/D and G (τ = -0.009, p= 0.98), and π/D and R (τ = -0.006, p= 0.89) in our combined neutralome dataset. All reported p-values are biased downwards as they fail to account for multiple testing and for linkage. This downward bias is conservative to our validation of neutrality. In contrast, repeating IRLS regression using all 872,398 nongenic loci yielded β_ν = 0.115, β_G =0.015 and a β_R = 0.024, with an overall regression F-statistic of 9,147 (all p<2.2e-16).

Estimating Ne_X / Ne_A

Ne_X/Ne_A estimated with π/D

Natural selection impacts absolute levels of diversity across the majority of the genome (Fig. 1). Differences in the strength of selection on the X chromosome relative to the autosomes at comparable genomic sites, such as regions near genes, also biases estimates of Ne_X/Ne_A. We estimated the effective population size on the X chromosome and the autosomes considering nongenic sites, sites far from genes (>0.3 cM), and sites in the neutralomes (Table 2) using π/D. We also estimated the ratio of the effective population sizes of the X chromosome (Ne_X) relative to the autosomes (Ne_A), as the ratio of median values for π/D (Table 3, Fig. 5). For nongenic loci, we infer that Ne_X/Ne_A is significantly smaller than 0.75 in gorillas, while all other taxa show the opposite trend (p < 0.001 across all comparisons). Information from regions far from genes inflates Ne_X/Ne_A, especially in the Yoruba and Nigerian chimpanzees (Table 2), causing Ne_X/Ne_A > 0.75 in all taxa. This is consistent with the (relatively) larger ν due to the reduced effects of linkage to *phastCons* elements on the X chromosome in regions far from genes (Fig. 4, Table S4). The estimate of ~1.0 in Yoruba is higher than the estimate from Gottipati et al.

(2011) (~0.91 when the same outgroup, orangutan, is used) taken at a similar distance to genes. When we bin the YRI data in accordance to the farthest bin of Gottipati et al. (2011) (0.2-0.4 cM), estimates of Ne_X/Ne_A (0.902, 95% CI: 0.875-0.938) become more similar. In humans, Ne_X/Ne_A estimate of 0.890 taken from the neutralome is quite close to the estimates of 0.882 based on genetic map lengths (Lohmueller et al. 2010),. Ne_X/Ne_A in the all taxa save bonobo exceed 0.75, though the neutralomes' relatively small size results in large standard errors. Ne_X/Ne_A in regions far from genes is significantly higher than the Ne_X/Ne_A inferred in the neutralomes in the YRI ($p < 0.001$, nonparametric bootstrap), while gorillas were only marginally significant at the 10% level ($p = 0.094$) and chimpanzees and bonobos were not significantly different ($p = 0.672$ and $p = 0.117$, respectively). When considering the neutralome, Ne_X/Ne_A is significantly smaller in bonobos than in chimpanzees ($p < 0.001$, nonparametric bootstrap).

Ne_X/Ne_A estimated with LD

To create a largely independent check on our inferences on N_f/N_m in the neutralome we estimated Ne_X/Ne_A using genetic maps as per Lohmueller et al. (2010) (SI Materials and Methods). LD-based genetic maps estimate the population genetic parameter ρ, which is equal to $4Ner$ where r is the rate of recombination measured in Morgans. We estimated Ne_X/Ne_A using ρ calibrated with r estimated in human pedigrees (Kong et al. 2010). Ne_X/Ne_A in the YRI was taken from Lohmueller et al. (2010). Ne_X/Ne_A estimated using genetic maps and the neutralome are consistent (Fig. 5, Table 3), which serves as a partial validation of the use of neutralomes to estimate Ne. Further, Ne_X/Ne_A is significantly smaller in bonobos than in chimpanzees ($p < 0.001$, nonparametric bootstrap) and significantly larger than 0.75 in all taxa ($p < 0.001$, nonparametric bootstrap).

DISCUSSION

The nongenic landscape of linked selection in great apes

Using fine-scale measurements on diversity and linkage across the genomes of great apes we show that both linkage to genes (*G*), as well as the local recombination rate (*R*), play a profound role in shaping genomic patterns of diversity in African great apes. Correlations between diversity and *G* and *R* are consistent with the action of selection at linked genic sites (to which *G* is sensitive) as well as nongenic sites (to which *R* is sensitive) (Maynard Smith and Haigh 1974; Charlesworth et al. 1993). Further, these correlations are no longer significant when we consider regions generally unlinked to *phastCons* elements (high ν) and far from genes (i.e. in neutralomes), demonstrating that genic and *phastCons* sites (and sites they are linked to) cause genome-wide reductions in diversity of not just humans (Woerner et al. 2016), but of African great apes. The effect-sizes of *G* and *R* vary across species and across chromosomes, and, while the effect size of linked selection to genes on the X chromosome is the largest amongst humans, gorillas and chimpanzees (Fig. 2), in bonobos the effect of linkage to genes on the X chromosome is approximately equal to that of the autosomes. Likewise, the effect size of linkage to genes on the autosomes (β_{AG}) is the smallest in bonobos, chimpanzees and humans, but not gorillas (Fig. 2). Contrasted to this, gorillas and bonobos have the highest and second-highest autosomal dN/dS ratios, respectively, amongst the apes studied here (Prado-Martinez et al. 2013), implying a paradoxical relationship between the strength of selection (*s*) (in terms of Nes) and the amount of the genome impacted by linked selection (in terms of β). This paradox is consistent with the expectations of weak selection, either of just negative selection (Williamson and Orive 2002), or of weak positive and negative selection (Tachida 2000). Both models predict that the strength of linked selection is strongest when |Nes| is neither too large nor too small, which may be the case on the autosomes in gorillas and not bonobos.

The neutralomes of great apes

The signals of demography and natural selection are easily conflated, where low levels of diversity can be consistent either with linkage to positively or negatively selected alleles (Maynard Smith and Haigh 1974; Charlesworth et al. 1993) or with small effective population size. We show that linked selection, specifically linked selection to genes and *phastCons* elements, reduces diversity throughout each of these genomes. Each taxon and each chromosome exhibits varying degrees of linked selection (Fig. 3), further complicating comparisons of Ne between species. We model the effects of linked selection, predicting the amount of diversity lost due to linkage to *phastCons* elements (our estimate ν) and extend the procedure of Woerner et al. (2016) by considering three types of *phastCons* elements to infer neutralomes in great apes. The neutralomes of nonhuman great apes are substantively smaller (~0.1% of genome) than that of humans (~1% of genome) (Table S5*B*). These size-differences may in part reflect issues of power. The genetic maps in non-human great apes were inferred from next-generation sequencing, thus the higher genotyping error rates that follow likely causes false evidence of recombination, making these maps more noisy than their human counterparts. The variable effects of linked genic selection also plays a substantive role in the sizes of the neutralomes. Taking gorillas for example, the selective effects of genes (Fig. 1, β_{AG} in Fig. 2, Table S1) caused much of the autosomes to be impacted by linked selection. Reflecting this, the autosomal neutralome in gorillas only uses regions at least ~0.4 cM from genes (b_G in Table S5), greatly reducing its size. Thus it follows that while the human neutralome is small by genomic, the neutralomes of non-human great apes are far smaller.

Estimates of Ne_X/Ne_A in great apes

We used two largely independent molecular estimates of Ne_X/Ne_A in great apes; estimates from diversity ratios, and estimates from genetic maps. Diversity-based estimate Ne_X/Ne_A are sensitive to differences in genic and nongenic selection between the X chromosome and the autosomes. Simply using regions far from genes does not appear to correct for variable levels of selection on the X chromosome

and the autosomes. While in most great apes the effects of linked selection are greater on the X chromosome (despite the reduced mutation rate and lower gene/*phastCons* densities of the X chromosome), regions far from genes instead show the opposite bias. Across African great apes we find that in regions far from genes the X chromosome has significantly higher ν (that is, less effects of linkage to *phastCons* elements) than the autosomes across apes (Fig. 5, Table S4), and roughly equal levels of diversity (Fig. S3). Differing ν far from genes indicates that linked selective pressures to *phastCons* elements changes from being more to less pronounced on the X chromosome, with the crossover point in ν being apparent, though variable, across great apes (Fig. 5). This in turn biases estimates of Ne_X/Ne_A. In particular, using just those regions "far" from genes (i.e., generally unlinked, though see b_G for gorillas and bonobos in Table S5), Ne_X/Ne_A is significantly higher (~1.0) in the Yoruba than in regions predicted to be neutral (~0.87). Further, the maximal *G* cutoff (0.4 cM) used by Hammer et al. (2010) and Gottipati et al. (2011) was both arbitrary and fortunate, as it appears to mitigate these variable nongenic selective effects. This maximum *G* cutoff is nearly identical to the minimum autosomal cutoff for *G* in neutralome of gorillas (table 3b), suggesting that this mitigation at 0.2-0.4 cM from genes is not universally applicable, but may instead be particular to inferences in humans. The use of neutralomes removes the need to be fortunate in our choice of thresholds, allowing for the direct inference of Ne_x/Ne_A, and greatly simplifying inferences on selective processes.

N_f/N_m in great apes

Mating and dispersal strategies are varied across great apes (Table 1), with these strategies leading to differences in effective population size on the X chromosome relative to the autosomes. While estimates of Ne_x/Ne_A may reflect differences in natural selection on the X chromosome relative to the autosomes, we present two estimates of N_f/N_m (that is, of the neutral Ne_x/Ne_A): the first is based on patterns of diversity in regions carefully selected to have no detectable levels of natural selection, while

the second is based on patterns of LD (Fig. 5; Table 3). These two measures are independent and coincident, with the confidence intervals in N_f/N_m from genetic maps falling strictly within the confidence intervals obtained from neutralomes (Table 3). Further, these two approaches are likely sensitive to measures of Ne at different time-scales (Lohmueller et al. 2010), suggesting that our findings are unlikely to be driven by changes in population size (e.g., that of Pool and Nielsen 2007). And, while other population dynamics such as sex-biased migration may contribute to our estimates of N_f/N_m, demographic effects other than higher variance in male reproductive success have at most a modest impact on estimates of N_f/N_m (Hammer et al. 2008).

Our estimates of N_f/N_m in great apes are consistent with previous works in humans (Gottipati et al. 2011; Arbiza et al. 2014) and bonobos (Prüfer et al. 2012). An estimate of ~2:1 in western lowland gorillas may appear surprising given their unimale mating system (Harcourt et al. 1981). However, if the harem structures in gorillas is unstable and have rapid turnover, then an estimate of 2:1 is not unexpected (Evans and Charlesworth 2013). Further subordinate male gorillas do have limited mating success (Watts 1990, 1991) which might contribute to our inference of N_f/N_m.

N_f/N_m in *Pan* is perhaps more surprising. Chimpanzees and bonobos have similar in mating structure and morphology which might otherwise suggest similar N_f/N_m in *Pan*. An N_f/N_m of ~2:1 is consistent with previous estimates in bonobos, and the ancestor of chimpanzees and bonobos (Table 3) (Prüfer et al. 2012). Our estimate of ~7-10:1 in Nigerian chimpanzees is perhaps unexpected large. Estimates of N_f/N_m across the four extant species of chimpanzees are very similar (Prado-Martinez et al. 2013), which makes demographic explanation for this finding less likely. Thus by parsimony, N_f/N_m may have increased within the lineage of *Pan troglodytes* and not in the ancestral lineage of *Pan*. Given this we posit two hypotheses that may explain these differences in N_f/N_m: The first hypothesis is that despite their similar testes size (Harcourt et al. 1981; Dixson 1998), sperm competition may vary considerably between chimpanzees and bonobos due to molecular, and not gross morphological, differences in these

taxa. Few studies have attempted to separate bonobos from chimpanzees in their molecular and morphological analyses (e.g., Anderson and Dixson 2002; Dorus et al. 2004; Nascimento 2008), and what few studies that have point to a limited amount of support for this hypothesis. In particular, Good et al. (2013) found modestly larger support for positive selection on ejaculated seminal fluid proteins in chimpanzees than in bonobos, though this finding may not be statistically significant. Kingan et al. (2003) also found significant evidence of positive selection between chimpanzees and humans at the semenogelin I locus, while similar tests between bonobos and humans were not significant. We note that our first hypothesis can either be viewed as an increase in sperm competition in *Pan troglodytes*, or an increase in the chimpanzee-bonobo ancestor and a separate loss in bonobos. Our second hypothesis is that both taxa have sperm competition, and that it is behavior that further elevates in N_f/N_m in chimpanzees. Chimpanzees practice considerably higher rates of infanticide (Wilson et al. 2014), and have mating that follows a more strict dominance hierarchy (Watts 1998; Gerloff et al. 1999; Furuichi 2011; Surbeck et al. 2011) than bonobos, while bonobos generally mate more than chimpanzees which may in turn reduce male-male competition (Furuichi 2011). Thus with our second hypothesis, behavioral tactics in combination with sperm competition may explain the elevated N_f/N_m in chimpanzees relative to bonobos. Ultimately, assays on the levels of sperm competition in bonobos compared to chimpanzees, as well as simulation approaches that model these behavioral differences, might serve to answer this outstanding evolutionary question in *Pan*.

METHODS

Samples

The samples used in this study come from two sources. The data in humans, the Yoruba (YRI) of Ibadan Nigeria, are from the publically available high coverage (~80x) genomes made available by Complete Genomics. The nonhuman data are a subsample of unrelated individuals from Prado-Martinez et al.

(2013), with most specimens being wild-born (though sampled in zoos). Sequence reads are accessible under the accession number SRP018689 (BioProject ID PRJNA189439) in NCBI's short read archive, with the processed SNP calls and masks available at: ftp://public_primate@biologiaevolutiva.org.

Mapping

The nonhuman species-specific read mappings and SNP-calling of Prado-Martinez et al. (2013) were used in all analyses. The YRI were mapped to hg19, the gorillas were mapped to gorGor3 (Ensembl 62), and both chimpanzees and bonobos were mapped to panTro4. Mapping details and filtering settings for the nonhuman samples are detailed in Prado-Martinez et al. (2013), while the filtering in the YRI is detailed in Woerner et al. (2016). In short, in the nonhuman samples, sites that passed variant quality score recalibration as well as had at least 7x coverage across all samples, and less than the 95th percentile coverage (summed over individual reads), as well as being more than 5 bp away from any indel, were included in our analysis. For the YRI we used the individual call-ability masks from the .tsv files from Complete Genomics, as well as removing sites within 5bp of any indel and sites that spanned copy-number or structural variants called by Complete Genomics. Genotype coordinates in gorillas were remapped from gorGor3 (Ensembl 62) to gorGor3.1 (Ensembl 72) using a custom script.

Gene definitions

Canonical protein-coding genes and the exons of canonical protein-coding genes were downloaded from Ensembl (build 72) for each taxon using an in-house perl script. As the number of genes and exons varied substantially across taxa (e.g. humans have 67.9 Mb of annotated canonical protein coding exon sequence, whilst gorillas have 46.4 Mb), we used annotations in humans to augment those that were species-specific. In particular, we used the UCSC liftOver tool to augment the species-specific genes with hg19's UCSC known genes (Downloaded on 2/6/2012). liftOver was run using the default parameters,

save the minMatch flag which was changed from the default of 0.95 to 0.50. Species-specific protein coding exon annotations were augmented in the same fashion, using instead human annotations from Ensembl build 72. All annotations within a species were merged using a base-wise union using a custom perl script.

Loci

While the approach of Woerner et al. (2016) used 10kb loci with a minimum of 1kb of callable sequence therein in their analysis, we instead chose to refine this approach for our regression analyses. Specifically, we carefully chose our loci to try to both maximize their number and to minimize their variance in the summary statistics computed upon them. To do this we constructed two types of masks; soft masks, which loci were allowed to overlap but within which no computations are to be performed, and hard masks, which no locus ever overlaps. We then specified two properties of a locus: the number of callable bases, which we set to 2kb, and the span of the locus, which was set at a maximum of 10kb. Thus each locus has exactly 2kb of bases in which computation is to be performed in a span (or length) of up to 10kb, and is guaranteed not to overlap any of the annotations in the hard mask. Given these constraints we set up a greedy program to extract the maximum number of such loci in each ape genome.

As our primary analysis seeks to quantify the effects of linkage to *phastCons* elements in the nongenic portions of genomes, we took for our hard mask the base-wise union of genes and *phastCons* elements inferred from primates and mammals, as well as the gaps in the species-specific genetic and physical maps. We also attempted to mask out *phastCons* clusters, but this resulted in a too substantial reduction in the number of loci. For soft masks we took the union of: sites not callable in either hg19 or rhesus (rheMac3) using the 100-way vertebrate alignments of UCSC, and places deemed uncallable in the resequencing data.

For our heatmap analyses, which are more qualitative and span genes, we also constructed a set of 10kb loci as per Woerner et al. (2016). Briefly, we partitioned the genome into nonoverlapping 10kb loci and excluded loci with less than 1kb of callable sequence. The same SNP-calling masks were used as with the 2kb loci, save the masking of genes.

Summary statistics

We computed nucleotide diversity (π) and divergence (D) to an inferred ancestral node in the tree. As our non-human data are lower coverage, use different sequencing technologies, and have a more modern processing pipeline than do the human data, we used slightly different processing to reflect these differences. In non-humans we induced haploid calling in males on the X chromosome by choosing whichever allele had the greatest number of reads supporting it, with ties being broken arbitrarily. In humans, we used the ploidy information provided in the SNP-calling by Complete Genomics. For non-humans, the ingroup base used in the D calculation was drawn uniformly at random from the alleles called at a site in the case that the site was segregating, otherwise the reference base was chosen. The outgroups used to polarize sites--rhesus (rheMac3) and humans (hg19)--were taken from the 100-way vertebrate alignments of UCSC for our nonhuman samples, while in humans we used the 46-way vertebrate alignments, polarizing sites using rhesus (rheMac2) and orangutan (ponAbe2). We computed D along the branch that leads to each species polarizing sites based on the allele called in rhesus.

Local estimates of Ne over sets of loci were estimated using the median π/D, with nonparametric bootstraps used to estimate the 95% CI. Estimates of the ratio Ne_X/Ne_A were computed taking the bootstrapped ratio of medians, with estimates of the standard error of the median taken as the standard deviation of the bootstrapped ratios. All bootstraps used 1000 iterations unless otherwise noted. Smoothing of π/D was done in the statistical package R. Generalized additive model smoothing was performed using the mgcv library in the geom_smooth function in ggplot2 using the default parameters.

Multiple IRLS Regression

With our 2kb loci we used IRLS linear regression to evaluate separately the effects of *R* and *G* on patterns of nongenic diversity. Specifically, using the R statistical package MASS, we modeled $\pi/D \sim G + R$ for both the autosomes and the X chromosome using the rlm function set to its default parameters. Standardized (β) coefficients, confidence intervals and slope significance were determined using a 1000-iteration nonparametric bootstrap as per Woerner et al. (2016). Visualization of the coefficients were performed in the R programming language using the coefplot function in the arm library.

DATA ACCESS

The code used to generate our estimates of ν across the genome, as well as the location all neutralomes, are available at: http://hammerlab.biosci.arizona.edu/Neutralome/neutralome.html

ACKNOWLEDGMENTS

Support for this work was provided by the US National Institutes of Health to M.F.H. (R01_HG005226) and NSF Graduate Research grant (DGE-1143953) to A.E.W.

DISCLOSURE DECLARATION

None

FIGURES

Figure 1. Heatmap of the median π/D (x 100) in four great apes. Heatmaps are shown from left to right for bonobos, Nigerian chimpanzees, a sub-Saharan African population in humans (YRI), and western lowland gorillas on the autosomes (top) and the X chromosome (bottom). Within each plot, the x-axis is binned by percentiles in the minimum genetic distance to genes and the y-axis by percentiles in the local recombination rate, with the first columns (G) being loci in genes. Darker cells correspond to higher median π/D within each plot, with the median values displayed in plain text.

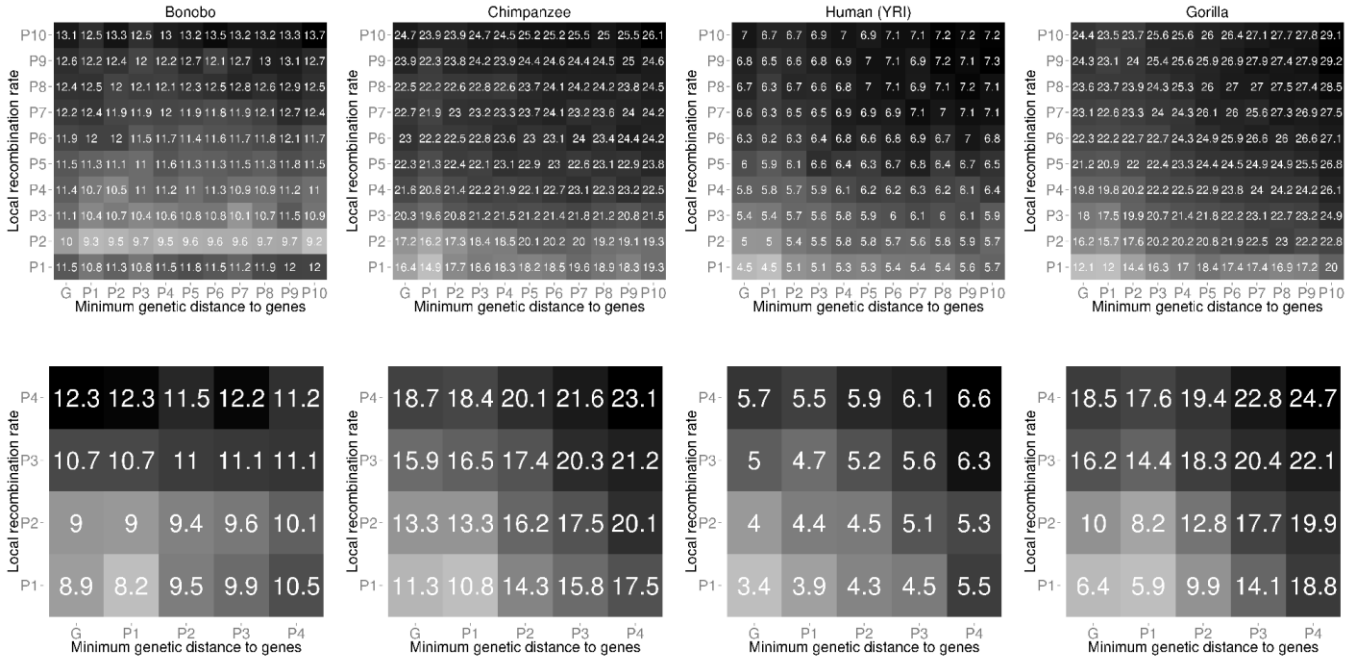


Figure 2. Estimates of the effect size (β) of linkage to genes (*G*) and the local recombination rate (*R*) on nongenic diversity. IRLS regression was used to $\pi/D \sim R + G$ separately on the autosomes and the X chromosome across four great apes, with the β-coefficients measuring the independent effect size for both variables. The coefficient estimates are shown ± their 95% CI across four apes (top to bottom).

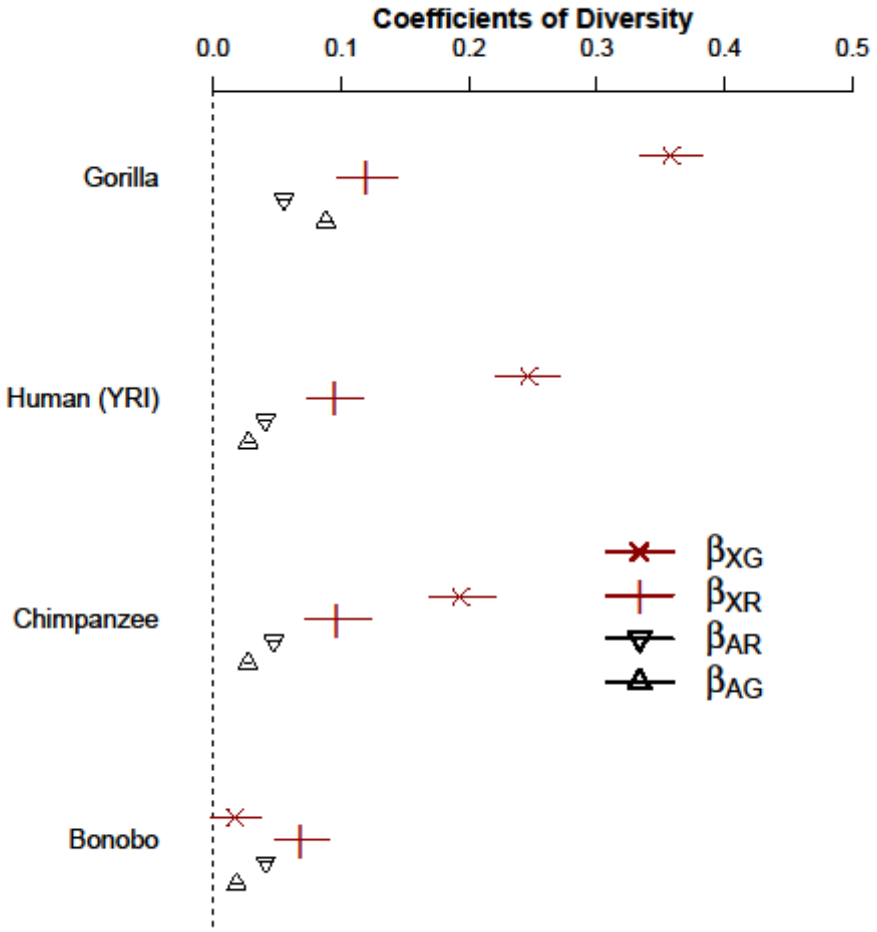


Figure 3. The empirical cumulative distribution function (ECDF) of the predicted nucleotide diversity relative to a neutral rate of 1 across four ape taxa. Predictions are based off of nonlinear least squares fitting of observed diversity levels to linkage to *phastCons* elements.

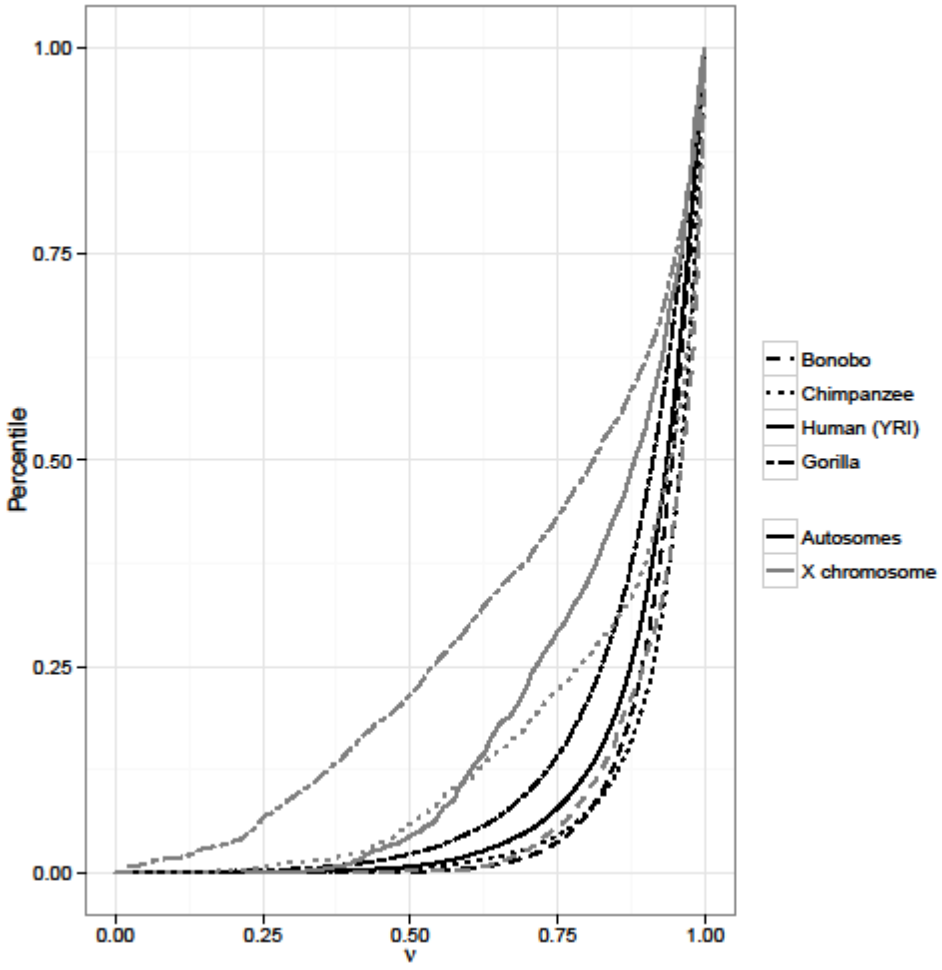


Figure 4. Predicted π/D (ν) versus the minimum genetic distance to genes. ν is given as a function of the minimum genetic distance to genes (G) on the X chromosomes (red dots) and the autosomes (blue dots). Note that both axes are truncated to highlight the differences between the X chromosome and the autosomes. ν was smoothed by a general additive model as a function of G on the X chromosome (dashed gray line) and the autosomes (solid black line). ν is given relative to a neutral rate of 1.0, and it is a prediction on the amount of diversity lost from phylogenetically conserved *phastCons* elements; a prediction made without explicit knowledge of genes. Smaller ν corresponds to greater predicted losses in diversity due to linked selection. On the X chromosome, regions near genes have greater predicted reductions in diversity than on the autosomes, while after the grey/black lines cross the opposite is true. Note that the *y*-axis is truncated to highlight differences between the X chromosome and the autosomes.

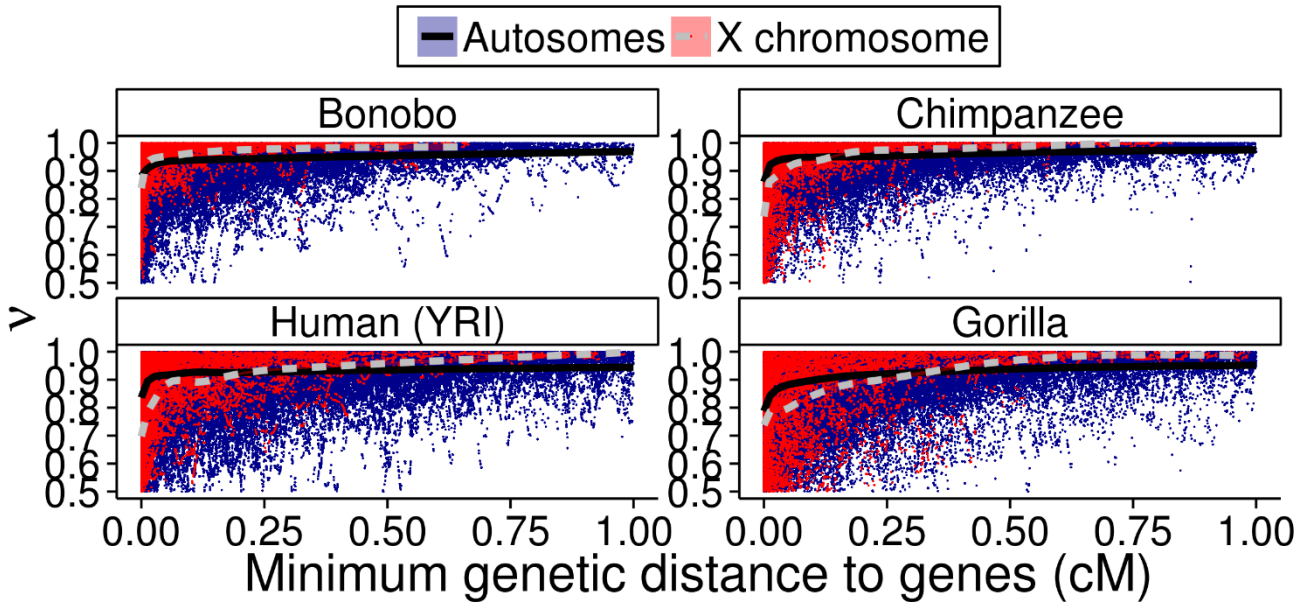


Figure 5. Estimates of the relative effective population size of the X chromosome (Ne_x) versus that of the autosomes (Ne_a) considering different assay types. Ne_x / Ne_a was estimated using the median π/D across all nongenic sites, just sites far from genes (>0.3 cM), and in the neutralome in four ape taxa. Ne_x / Ne_a was also estimated using patterns of linkage disequilibrium in genetic maps. 95% confidence intervals were estimated using a 1000-iteration bootstrap.

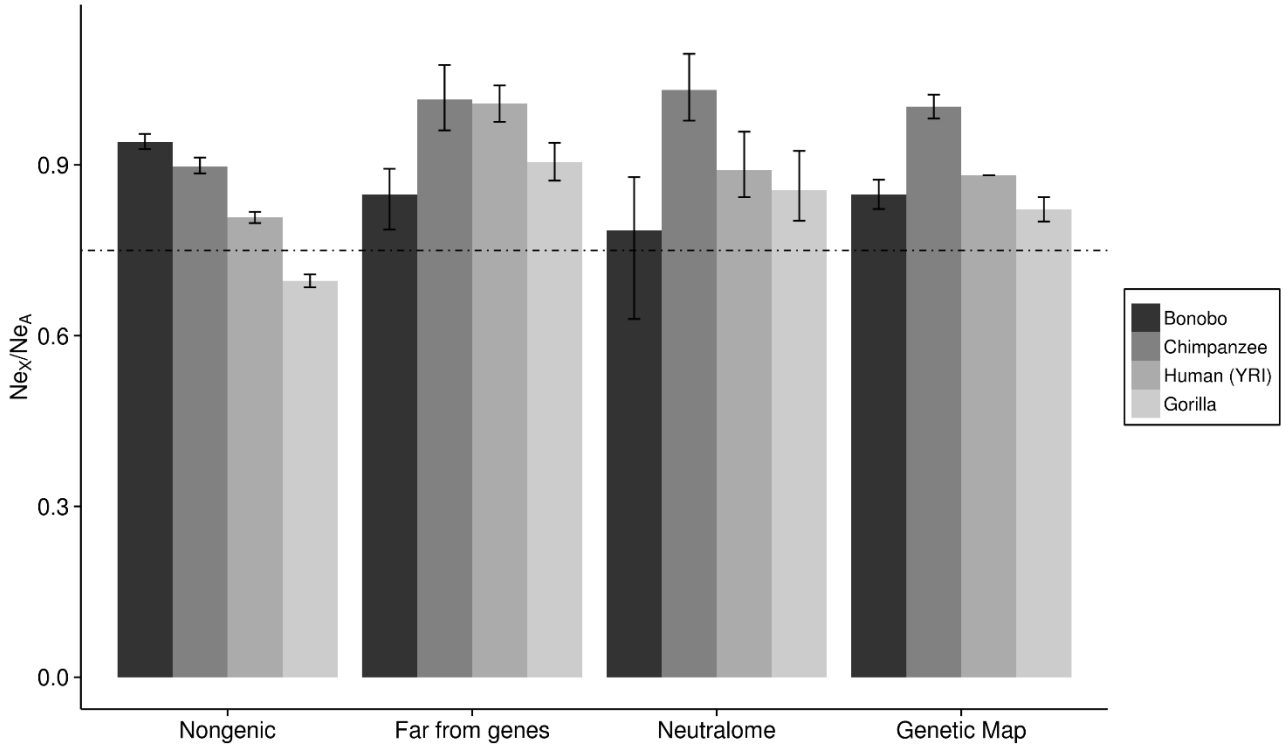


TABLE LEGENDS

Table 1. Mating structures in nonhuman great apes. Chimpanzees and bonobos have multimale-multifemale (MM) mating systems, while gorillas have a polygynous unimale (UM-P) system. Body weight dimorphism (male/female body weight) can be extreme (>2), strong (1.5-1.9) or moderate (1.1-1.4) (Dixson 1998). Testes/body weight ratios (g/kg) can be small (< 0.5), moderate (0.5-1.5) or large (>2.0) (Harcourt et al. 1981).

Table 2. π/D on the X chromosome and the autosomes across apes according to different samples of the genome. π/D (medians) are shown, as well as standard errors of the median, for all nongenic regions, regions far from genes (> 0.3 cM), and the neutralome. Standard errors were determined with a 1000-iteration bootstrap.

Table 3. Estimates of Ne_X/Ne_A and N_f/N_m across apes according to different assays of the genome. Diversity-based Ne_X/Ne_A was estimated by taking the ratio of the median π/D of the X chromosome versus the autosomes using nongenic regions, regions far (>0.3cM) from genes, and in the neutralomes. Genetic map estimates of Ne_X/Ne_A were computed using Ne estimated in genetic maps calibrated by pedigrees. Confidence intervals were obtained by a 1000-iteration nonparametric bootstrap, as well as the ability to reject a 1:1 sex ratio (Ne_X/Ne_A =0.75). Ne_X/Ne_A values were directly converted into α values using standard population genetic theory. * Ne_X/Ne_A in the YRI is taken from Lohmueller et al. (2010). ** Ne_X/Ne_A estimated in the ancestor of *Pan* using incomplete lineage sorting (ILS) was taken from Prüfer et al. (2012).

TABLES

Table 1. Mating structures in nonhuman great apes.

Taxon	Mating system	Body weight dimorphism	Testes/body weight ratio
Pan troglodytes	MM	Moderate	Large
Pan paniscus	MM	Moderate	Large
Gorilla gorilla	UM-P	Extreme	Small

Table 2. Median π/D and standard errors of the median considering different classes of sites.

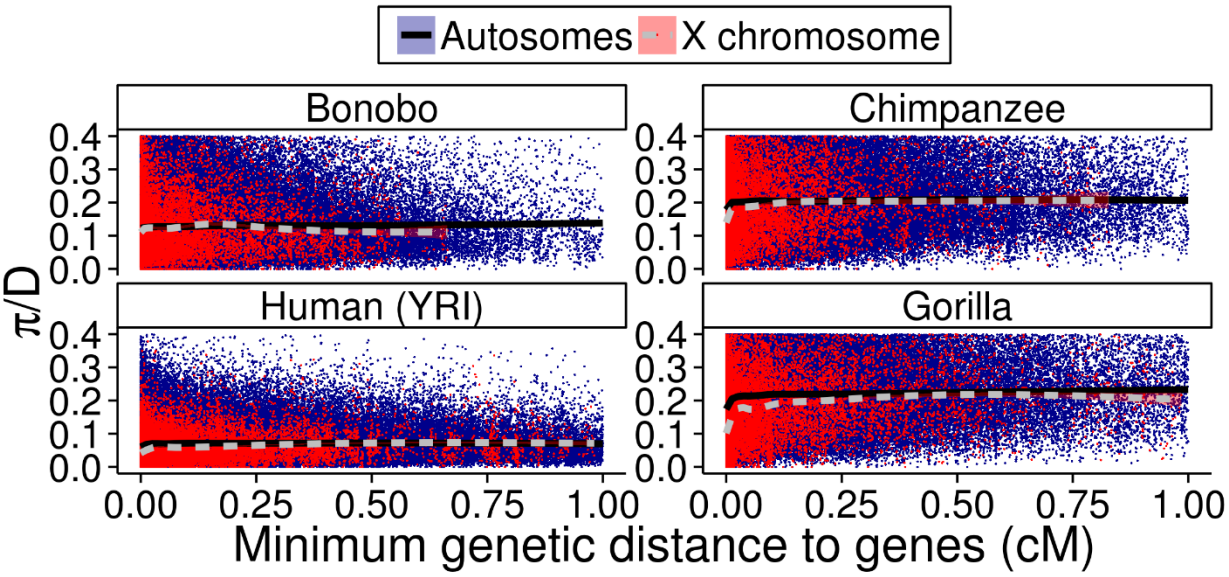
Taxon	**Regions considered**	**X chromosome**	**SE**	**Autosomes**	**SE**
Bonobo	Nongenic	0.106	6.96E-04	0.113	2.30E-04
Bonobo	Far from genes	0.099	2.87E-03	0.117	6.85E-04
Bonobo	Neutralome	0.117	2.98E-03	0.149	1.08E-02
Chimpanzee	*Nongenic*	0.196	1.57E-03	0.218	3.93E-04
Chimpanzee	*Far from genes*	0.232	7.17E-03	0.229	1.06E-03
Chimpanzee	*Neutralome*	0.258	5.27E-03	0.250	4.82E-03
Human (YRI)	Nongenic	0.048	3.01E-04	0.060	9.68E-05
Human (YRI)	Far from genes	0.062	1.02E-03	0.062	2.15E-04
Human (YRI)	Neutralome	0.062	2.05E-03	0.070	4.15E-04
Gorilla	*Nongenic*	*0.164*	*1.37E-03*	*0.235*	*4.27E-04*
Gorilla	*Far from genes*	*0.236*	*4.18E-03*	*0.261*	*9.97E-04*
Gorilla	*Neutralome*	*0.256*	*8.30E-03*	*0.299*	*4.56E-03*

Table 3. Estimates of Ne_X/Ne_A and Nf/Nm in hominoids.

Taxon	Assay Type	Ne_X/Ne_A	CI: 0.025	CI: 0.975	*P(Nex/Nea) = 0.75*	Nf/Nm (expectation)	CI: 0.025	CI: 0.975
Bonobo	*π/D* (Nongenic)	0.94	0.93	0.95	< 0.001	4.08	3.70	4.60
Chimp	*π/D* (Nongenic)	0.90	0.89	0.91	< 0.001	2.94	2.69	3.31
Gorilla	*π/D* (Nongenic)	0.70	0.68	0.71	< 0.001	0.63	0.56	0.69
Human (YRI)	*π/D* (Nongenic)	0.81	0.80	0.82	< 0.001	1.55	1.44	1.66
Bonobo	*π/D* (Far from genes)	0.85	0.79	0.89	< 0.001	2.06	1.32	2.86
Chimp	*π/D* (Far from genes)	1.02	0.96	1.08	< 0.001	8.23	4.85	20.74
Gorilla	*π/D* (Far from genes)	0.90	0.87	0.94	< 0.001	3.10	2.45	4.04
Human (YRI)	*π/D* (Far from genes)	1.01	0.98	1.04	< 0.001	7.64	5.54	11.18
Bonobo	*π/D (Neutralome)*	*0.78*	*0.63*	*0.88*	*0.45*	*1.31*	*0.27*	*2.56*
Chimp	*π/D (Neutralome)*	*1.03*	*0.98*	*1.10*	*< 0.001*	*10.08*	*5.64*	*35.55*
Gorilla	*π/D (Neutralome)*	*0.86*	*0.80*	*0.92*	*0.002*	*2.18*	*1.48*	*3.61*
Human (YRI)	*π/D (Neutralome)*	*0.89*	*0.84*	*0.96*	*< 0.001*	*2.80*	*1.99*	*4.75*
Bonobo	*Genetic Map*	*0.85*	*0.82*	*0.87*	*< 0.001*	*2.06*	*1.72*	*2.48*
Chimp	*Genetic Map*	*1.00*	*0.98*	*1.02*	*< 0.001*	*7.19*	*5.83*	*9.08*
Gorilla	*Genetic Map*	*0.82*	*0.80*	*0.84*	*< 0.001*	*1.70*	*1.47*	*1.99*
Human (YRI)[*]	*Genetic Map*	*0.88*	--	--		*2.63*	--	--
Pan ancestor[**]	*ILS*	0.83	0.75	0.91		*1.81*	*1.00*	*3.23*

SUPPLEMENTAL FIGURES

Supplemental Figure 1. π/D versus the minimum genetic distance to genes. π/D is given as a function of the minimum genetic distance to genes (G) on the X chromosomes (red dots) and the autosomes (blue dots). π/D was smoothed by a general additive model as a function of G on the X chromosome (dashed gray line) and the autosomes (solid black line). Unlike ν, π/D is given as an absolute value; nevertheless π/D on the X chromosome exceeds that of the autosomes in some taxa with loci far from genes. Note that both axes were truncated so as to highlight differences between the X chromosome and the autosomes.



SUPPLEMENTAL TABLES

Supplemental Table 1. Beta-coefficient estimates from the IRLS regression modeling $\pi/D \sim R + G$, where R and G are the local recombination and the minimum genetic distance to genes, respectively. Separate regressions were run on the X chromosome (X) and the autosomes (A).The coefficient point estimates are given, plus the 95% confidence intervals. Also given is the probability that the coefficient is positive, and the probability that the coefficient for a given row is greater than the row that follows.

Taxon	Coefficeint	Point Estimate	CI: 0.025	CI: 0.975	*Pr coef > 0*	Pr this coef > next coef
Bonobo	β_{XR}	0.068	0.048	0.090	< 0.001	0.006
Bonobo	β_{AR}	0.041	0.038	0.045	< 0.001	0.011
Bonobo	β_{XG}	*0.017*	*-0.002*	*0.037*	*0.049*	*0.474*
Bonobo	β_{AG}	*0.018*	*0.016*	*0.021*	*< 0.001*	
Chimpanzee	β_{XG}	*0.194*	*0.168*	*0.220*	*< 0.001*	*< 0.001*
Chimpanzee	β_{XR}	0.096	0.071	0.124	< 0.001	< 0.001
Chimpanzee	β_{AR}	0.047	0.044	0.051	< 0.001	< 0.001
Chimpanzee	β_{AG}	*0.027*	*0.024*	*0.030*	*< 0.001*	
Human (*YRI)*	β_{XG}	*0.247*	*0.222*	*0.271*	*< 0.001*	*< 0.001*
Human (YRI)	β_{XR}	0.094	0.073	0.117	< 0.001	< 0.001
Human (YRI)	β_{AR}	0.041	0.037	0.044	< 0.001	< 0.001
Human (YRI)	β_{AG}	*0.027*	*0.024*	*0.031*	*< 0.001*	
Gorilla	β_{XG}	*0.358*	*0.333*	*0.383*	*< 0.001*	*< 0.001*
Gorilla	β_{XR}	0.119	0.097	0.144	< 0.001	0.001
Gorilla	β_{AG}	*0.088*	*0.084*	*0.092*	*< 0.001*	*< 0.001*
Gorilla	β_{AR}	0.055	0.051	0.060	< 0.001	

Supplemental Table 2. Estimates of model fit for Model C of Halligan et al. (2013). The expected π/D was computed given linkage to different classes of phylogenetically conserved *phastCons* elements (Element) on both the autosomes (A) and the X chromosome (X) across four ape taxa. π/D was fit using a sum of squares criterion (SS) to both the model of Halligan et al. (2013) contrasted to that of the mean π/D (SS Total). Levels of model fit (measured as R^2) are shown, with the highest model fits being highlighted in grey.

Taxon	Compartment	Element	SS Model	SS Total	R^2
Bonobo	A	Primates	3,237.144	3,259.562	0.69%
Bonobo	X	Primates	172.067	173.932	1.07%
Bonobo	A	Mammal	3,237.297	3,262.157	0.76%
Bonobo	X	Mammal	172.413	174.155	1.00%
Bonobo	A	Primate Clusters	3,235.715	3,262.148	0.81%
Bonobo	X	Primate Clusters	172.093	174.155	1.18%
Chimpanzee	A	Primates	7,341.674	7,433.626	1.24%
Chimpanzee	X	Primates	437.088	462.631	5.52%
Chimpanzee	A	Mammal	7,328.056	7,434.994	1.44%
Chimpanzee	X	Mammal	438.880	465.316	5.68%
Chimpanzee	A	Primate Clusters	7,323.968	7,434.588	1.49%
Chimpanzee	X	Primate Clusters	438.092	465.316	5.85%
Human (YRI)	A	Primates	502.691	515.137	2.42%
Human (YRI)	X	Primates	31.266	33.425	6.46%
Human (YRI)	A	Mammal	501.882	515.634	2.67%
Human (YRI)	X	Mammal	31.362	33.584	6.61%
Human (YRI)	A	Primate Clusters	501.534	515.174	2.65%
Human (YRI)	X	Primate Clusters	31.383	33.604	6.61%
Gorilla	A	Primates	4,510.975	4,733.455	4.70%
Gorilla	X	Primates	227.895	284.192	19.81%
Gorilla	A	Mammal	4,495.563	4,735.844	5.07%
Gorilla	X	Mammal	226.290	284.192	20.37%
Gorilla	A	Primate Clusters	4,489.587	4,738.314	5.25%
Gorilla	X	Primate Clusters	225.961	284.192	20.49%

Supplemental Table 3A. Mean nongenic ν on the autosomes and the X chromosome, and their difference, after matching the recombination rate on the autosomes and the X chromosome. Each nongenic locus on the X chromosome was paired to an autosomal locus on the basis of its local recombination rate, and a paired 2-sample *t*-test was used to assess the difference in means

Taxon	Autosomal mean ν	X chromosome mean ν	Mean of the difference	0.025 CI	0.975 CI	*p*-value
Bonobo	0.916	0.927	-0.011	-0.013	-0.009	< 2.2e-16
Chimp	0.916	0.860	0.056	0.052	0.059	< 2.2e-16
Human (YRI)	0.884	0.825	0.059	0.057	0.062	< 2.2e-16
Gorilla	0.851	0.725	0.126	0.121	0.131	< 2.2e-16

Supplemental Table 3B. Mean nongenic ν on the autosomes and the X chromosome. An unpaired Welch *t*-test was used to test the difference in means, of which the p-value and 95% confidence interval are shown

Taxon	Autosomal mean ν	X chromosome mean ν	0.025 CI	0.975 CI	*p*-value
Bonobo	0.911	0.915	-0.005	-0.002	0.0001464
Chimp	0.919	0.845	0.071	0.078	< 2.2e-16
Human (YRI)	0.890	0.808	0.079	0.084	< 2.2e-16
Gorilla	0.852	0.698	0.149	0.158	< 2.2e-16

Supplemental Table 4A. Mean nongenic ν on the autosomes and the X chromosome, and their difference, after matching the recombination rate on the autosomes and the X chromosome in regions far (> 0.3cM) from genes. Each nongenic locus on the X chromosome was paired to an autosomal locus on the basis of its local recombination rate, and a paired 2-sample t-test was used to assess the difference in means

Taxon	Autosomal mean ν	X chromosome mean ν	Mean of the difference	0.025 CI	0.975 CI	*p*-value
Bonobo	0.953	0.981	-0.028	-0.033	-0.024	< 2.2e-16
Chimp	0.966	0.988	-0.021	-0.024	-0.019	< 2.2e-16
Human (YRI)	0.937	0.960	-0.023	-0.026	-0.019	< 2.2e-16
Gorilla	0.936	0.955	-0.019	-0.024	-0.015	1.23E-015

Supplemental Table 4B. Mean nongenic ν on the autosomes and the X chromosome in regions far (>0.3cM) from genes. An unpaired Welch *t*-test was used to test the difference in means, of which the *p*-value and 95% confidence interval are shown

Taxon	Autosomal mean ν	X chromosome mean ν	0.025 CI	0.975 CI	*p*-value
Bonobo	0.953	0.981	-0.031	-0.025	< 2.2e-16
Chimp	0.964	0.988	-0.026	-0.022	< 2.2e-16
Human (YRI)	0.936	0.960	-0.025	-0.021	< 2.2e-16
Gorilla	0.937	0.955	-0.021	-0.014	< 2.2e-16

Supplemental Table 5A. The inference of neutralomes in several species of apes conditioning on ν. Using our iterative breakpoint analysis we inferred the location of neutralomes based on a threshold (breakpoint) in ν separately for the autosomes and the X chromosome, and separately for each species. Breakpoints in ν were inferred using an iterative procedure, where weighted nonlinear regression was used to propose a breakpoint that demarkates the neutralome (Nu > bν). In each iteration of the procedure, loci in the proposed neutralome are tested for significant slopes using multiple linear regression, and if the minimum p-value is > 0.10 the procedure stops, otherwise it continues using all loci after the breakpoint. If 50 or fewer loci were detected in the proposed neutralome the procedure halts and is considered to have failed. The penultimate bν shows the previous value of ν that failed to separate the neutralome from the non-neutralome.

Taxon	Compartment	Ultimate b_ν	Min *p*-value (uncorrected)	Number of Loci	Number of iterations	Penultimate b_ν
Bonobo	Autosomes	--	--	< 50	1	98.500%
Bonobo	X	99.786%	0.142	137	2	99.386%
Chimp	Autosomes	99.876%	0.121	1,774	3	99.166%
Chimp	X	99.976%	0.103	122	4	99.936%
Human (YRI)	Autosomes	99.582%	0.435	5,438	1	98.500%
Human (YRI)	X	99.405%	0.589	251	1	98.500%
Gorilla	Autosomes	--	--	< 50	5	99.904%
Gorilla	X	99.655%	0.173	339	1	98.500%

Supplemental Table 5B. The inference of neutralomes in several species of apes conditioning on ν and G. Using our iterative breakpoint analysis, we inferred the location of neutralomes based on a threshold (breakpoint) in ν and the genetic distance to genes (G) separately for the autosomes and the X chromosome, and separately for each taxon. Breakpoints in ν and G were inferred using an iterative procedure, where weighted nonlinear regression was used to propose a breakpoint that demarkates the neutralome. In each iteration of the procedure, loci in the proposed neutralome are tested for significant slopes using multiple linear regression, and if the minimum p-value is > 0.10 the procedure stops, otherwise it continues using all loci after the breakpoint.

Taxon	Compartment	Ultimate b_ν	Ultimate b_G	Min *p*-value (uncorrected)	Number of Loci	Number of Iterations	Penultimate b_ν	Penultimate b_G
Bonobo	Autosomes	99.842%	0.398	0.715	118	6	99.584%	0.281
Bonobo	X	98.757%	0.106	0.157	1,260	1	98.500%	0.000
Chimp	Autosomes	99.700%	0.217	0.149	1,987	5	98.669%	0.143
Chimp	X	98.726%	0.015	0.225	1,663	1	98.500%	0.000
Human (YRI)	Autosomes	98.509%	0.121	0.127	13,681	1	98.500%	0.000
Human (YRI)	X	98.990%	0.127	0.294	421	1	98.500%	0.000
Gorilla	Autosomes	99.140%	0.399	0.318	1,398	4	99.097%	0.320
Gorilla	X	99.595%	0.158	0.391	260	1	98.500%	0.000

SI Material and Methods:

Genetic maps

For the nonhuman great apes we used the species-specific genetic maps from Stevison et al. (2016) to compute levels of linkage, while for humans we used the population-averaged (CEU+YRI) HapMap genetic map (Frazer et al. 2007). We also generated genetic maps for the X chromosome in nonhuman great apes using the framework described in Stevison et al. (2016). While the HapMap genetic map is relatively complete, the genetic maps in nonhuman great apes were computed solely within regions that are syntenic across great apes. While most of the genome is covered using this approach (~2.6 Gb), gaps in the genetic map are problematic for computing linkage between distant sites on the chromosomes (e.g. between a locus and the nearest gene). To address this we filled the gaps between neighboring syntenic regions using the chromosome-averaged ρ (that is, 4N*e*r, where *r* is the rate of recombination in Morgans), and we then added these nonsyntenic gaps to our SNP calling mask (see Methods). This effectively removed nonsyntenic loci from our analysis whilst retaining information on linkage. Note that genes and conserved elements that occur in these gaps were, however, retained in our linkage analyses.

We estimated Ne_X/Ne_A with genetic maps using the approach of Lohmueller et al. (2010). Briefly, we used ρ inferred from LD-based genetic maps estimate Ne. To infer Ne using genetic maps one needs a direct measure of *r*, with this value typically being inferred in pedigrees. To do this we measured the cumulative *r* across the syntenic regions of Stevison et al. (2016) in the decode genetic map in hg18 coordinates (Kong et al. 2010). Note we considered using *r* estimated in chimpanzees (Venn et al. 2014), however their estimates are especially coarse and rely on only 9 recombination events on the X chromosome, which would likely add too much variance to our estimates. Taking these same syntenic regions we also computed the cumulative ρ across the X chromosome and the autosomes in nonhuman great apes. Using these cumulative values we solved the ratio NeX/NeA and we estimated confidence

intervals using a 1000-iteration nonparametric bootstrap. NeX/NeA in the YRI was taken from Lohmueller et al. (2010).

Modeling genome-wide diversity levels

We downloaded *phastCons* conserved elements inferred from primates and placental mammals, as well as *phastCons* scores (as opposed to elements) inferred from primates from the UCSC genomic database. *phastCons* scores from primates were naïve clustered by taking just those base-positions whose scores were most conserved, and clustering contiguous positions into elements. To control the number of elements created this way we attempted to match the number of bases in primate elements to the number of bases in our naïve clustering. Using a counting sort on the fixed-precision *phastCons* score, we found the minimum *phastCons* score such that at most the number of bases in the primate elements would be found. Conserved elements were mapped from the hg19 genome to the species-specific physical maps using UCSC's liftOver tool with the min-match parameter set to 0.5.

We applied "Model C" of Halligan et al. (2013) to model the reductions in diversity that stem from linkage to *phastCons* elements in Woerner et al. (2016). We considered two classes of elements: *phastCons* elements that either intersected protein-coding regions (conserved exonic elements, CEEs) or those that did not (conserved nonexonic elements, CNEs). CEE and CNE elements were defined using bedtools intersect (Quinlan and Hall, 2010) for each of the three types of *phastCons* element described above and for each species using the species-specific exon annotations.

In Model C, π/D decreases exponentially near selective elements, with the predicted diversity at a given nonselective site in the genome being the product over all linked sites. Diversity at a single site is approximated by:

$$\frac{\pi}{D} \sim \exp\left(\log(p_1) - p_2 \sum e^{-\frac{x_i}{p_3}} - p_4 \sum e^{-\frac{x_i}{p_5}}\right)$$

Where p_1 is the neutral or unreduced level of diversity, p_2 and p_3 are the reductions in diversity observed at a CEE and CNE sites, respectively, and p_2 and p_4 are the rates of decay around CEE and CNE sites, respectively. x_i is the distance in morgans $(M)^{-8}$ from the nonselective site to the ith selective site.

Using the approach of Woerner et al. (2016) we computed the expected π/D in the center of each 2kb locus considering all *phastCons* elements within 1 cM of said locus. Parameters p_1–p_5 were estimated under a sum of squares criterion using Nelder-Mead optimization in a random-sampling framework as per Woerner et al. (2016). Briefly, for the X chromosome 10,000 random starting points in our 5-dimensional parameter space were chosen, and Nelder Mead optimization was used on each of them to find a local minimum. As the autosomes is ~20x the size of the X chromosome, we randomly sampled 1/20 loci on the autosomes, and used 10,000 random starting points as per the X chromosome. The top 1,000 discrete points were taken from this and chosen as starting points in the optimization procedure for the whole of the autosomes. This procedure was applied separately for all taxa, and separately for all three of our classes of *phastCons* elements, and we report on the best fitting models (Table S1). The final result of this parameter estimation is the expected relative effective population size, ν, for each of our nongenic 2kb loci across all taxa, as well as the coefficient of determination (R^2) (Table S1). ν was estimated independently for each of the three classes of *phastCons* elements.

Constrained Segmented Regression

Using our estimates of ν we set out to identify regions of the genome in which the effects of linked selection are minimal, i.e., neutralomes, as defined in Woerner et al. (2016). Briefly, while ν measures levels of linked selection to *phastCons* elements, other regions of the genome are also likely under selection. If we make the simplifying assumption that other selective elements exist on a background that contains some appreciable density of *phastCons* elements, then it follows that being unlinked to *phastCons* elements may become sufficient to be unlinked to any selective site. Given this, we reduced the problem of finding a neutralome to the problem of finding a breakpoint in a constrained segmented

(a.k.a. broken stick) regression. Specifically, we considered the relationship between π/D and ν as having two phases; a positively correlated phase, and a second phase where π/D plateaus (i.e. a connected line segment with no slope) when ν is sufficiently large. This can be modeled with three parameters: two parameters for the intercept and the slope in the first phase, and another parameter for the breakpoint. The predicted values of π/D in the plateau after the breakpoint coming from the correlated phase's linear predictions at the value of the breakpoint. Nonlinear optimization techniques can be used to solve for these three parameters, though this approach requires five parameters once we consider thresholding on the distance to genes (G) (below). Instead, we used a simple heuristic to model a 1-dimensional optimization problem parameterized solely on the breakpoint. Given a breakpoint b in ν, we can use IRLS regression to model $\pi/D \sim \nu$ on the loci in the correlated phase ($\nu \leq b$), and we can use the coefficients from the regression to form two line segments; the predictions from the first phase coming directly from the regression coefficients and the second phase's plateau predictions are just the predicted π/D values at the breakpoint. As finding b is just a 1-dimensional nonlinear optimization problem, we used Brent's method with the optim function in R to solve for b under a weighted sum of squares criterion. To accommodate outliers we use a weighted sum of squares criterion, using the weights inferred from the IRLS regression model that we use to assess neutrality (see below), i.e. $\pi/D \sim \nu + G + R$. This served to both reduce the effect of outliers as well as to make our object function in our assessment of neutrality and our search for neutrality more similar.

Expanding this search strategy to also consider G requires two additional parameters: a second slope for G in the correlated phase, and a second breakpoint in G to demarcate the plateau. Similar to the single breakpoint approach, given two breakpoints, b_ν and b_G, we use IRLS regression on loci with $\nu \leq b_\nu$ or $G \leq b_G$ to solve for the intercept and two slopes, as well as the expected value at the plateau. We then used Nelder-Mead optimization to solve for b_ν and b_G using the optim function in R under the same weighted sum of squares criterion as above. Given a solution for either one or two breakpoints, we can

then test to see if the loci in the plateau appear neutral as per Woerner et al. (2016). Namely, we modeled $\pi/D \sim \nu + G + R$, and we used the minimum 1-tailed *p*-value of the coefficients as an indicator of neutrality.

Inference of neutralomes

Now that we have established a technique to find breakpoints, we used the following iterative technique to identify neutralomes. As we have estimates of ν based on three different types of *phastCons* elements, we used as our estimator of ν the minimum ν (i.e., the worst-case ν) inferred across these element-types to ensure that we were largely unlinked to all types of *phastCons* elements. For the 1-breakpoint approach we chose an initial value of ν (98.5%) that we considered nearly neutral, and applied our segmented regression approach to solve for *b*. Considering only those loci with $\nu > b$, we considered the set of loci after the breakpoint to be neutral if the uncorrected minimum *p*-value overall all coefficients was > 0.1. Otherwise we iterated our procedure, using our segmented regression technique only on those loci with $\nu > b$ (Table S5*A*). This approach was applied separately to the autosomes and the X chromosome in each taxa. We applied the same iterative technique to our 2-breakpoint framework, starting with an initial *G* of 0, to define a second set of neutralomes (Table S5*B*).

REFERENCES

1. Anderson MJ, Dixson AF. 2002. Sperm competition: motility and the midpiece in primates. Nature. 416: 496.
2. Arbiza L, Gottipati S, Siepel A, Keinan A. 2014. Contrasting X-linked and autosomal diversity across 14 human populations. Am J Hum Genet 94:827-44.
3. Begun DJ, Whitley P. 2000. Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proc Natl Acad Sci 97:5960-5.
4. Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci **85**: 6414-8.
5. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 42: 806-10.

6. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 140:783-96.
7. Caballero A. 1995. On the effective size of populations with separate sexes, with particular reference to sex-linked genes. Genetics. 139: 1007-11.
8. Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am Nat. 130:113-46.
9. Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-303.
10. Charlesworth B. 1996. Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res. 68:131-50.
11. Charlesworth B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet Res 77: 153-66.
12. Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet 80: 692-704.
13. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-13.
14. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 6: e1001025.
15. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, et al. 1998. A physical map of 30,000 human genes. Science. 282: 744-6.
16. Dixson A. 1998. Primate Sexuality: Comparative Studies of the Prosimians, Monkeys, Apes, and Human Beings. New York, NY: Oxford University Press.
17. Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet. 36: 1326-9.
18. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. **327**: 78-81.
19. Emery LS, Felsenstein J, Akey JM. 2010. Estimators of the human effective sex ratio detect sex biases on different timescales. Am J Hum Genet. 87: 848-56.
20. Evans BJ, Charlesworth B. 2013. The effect of nonindependent mate pairing on the effective population size. Genetics. 193: 545-56.
21. Furuichi T. 2011. Female contributions to the peaceful nature of bonobo society. Evolutionary Anthropology: Issues, News, and Reviews. 20: 131-42.
22. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
23. Goldberg A, Rosenberg NA. 2015. Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture on the X chromosome. Genetics. 201: 263-79.
24. Good JM, Wiebe V, Albert FW, Burbano HA, Kircher M, Green RE, Halbwax M, André C, Atencia R, Fischer A, Pääbo S. 2013. Comparative population genomics of the ejaculate in humans and the great apes. Mol Biol Evol. 30: 964-76.
25. Gerloff U, Hartung B, Fruth B, Hohmann G, Tautz D. 1999. Intracommunity relationships, dispersal pattern and paternity success in a wild living community of Bonobos (Pan paniscus) determined from DNA analysis of faecal samples. Proceedings of the Royal Society of London B: Biological Sciences. 266: 1189-95.

26. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A. 2011. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. Nat Genet 43:741-3.
27. Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. Mol Biol Evol 28: 2651-60.
28. Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet 5;9:e1003995.
29. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet. 4: e1000202.
30. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD. 2010. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42: 830-831.
31. Harcourt AH, Harvey PH, Larson SG, Short RV. 1981. Testis weight, body weight and breeding system in primates. Nature. 293: 55-7.
32. Hedrick PW. 2007. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution. 61: 2750-71.
33. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. Science. 331: 920-4.
34. Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. Proc Natl Acad Sci 109:2054-9.
35. Kano T. The last ape: Pygmy chimpanzee behavior and ecology. Stanford: Stanford University Press; 1992
36. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. Science. 317: 915
37. Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41: 66-70.
38. Kingan SB, Tatar M, Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. Journal of Molecular Evolution. 57:159-69.
39. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, Gudjonsson SA. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 467: 1099-103.
40. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, igurdsson A, Jonasdottir A, Jonasdottir A et al. 2012. Rate of de novo mutations and the importance of father/'s age to disease risk. Nature 488: 471-475.
41. Labuda D, Lefebvre JF, Nadeau P, Roy-Gagnon MH. 2010. Female-to-male breeding ratio in modern humans—an analysis based on historical recombinations. Am J Hum Genet. 86: 353-63.
42. Laporte V, Charlesworth B. 2002. Effective population size and population subdivision in demographically structured populations. Genetics. 162: 501-19.
43. Lercher MJ, Urrutia AO, Hurst LD. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. Mol Biol Evol 20: 1113-6.
44. Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, et al.. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet 7: e1002326.
45. Lohmueller KE, Degenhardt JD, Keinan A. 2010. Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al. Am J Hum Genet 86(6):978-80.
46. Low BS. 1988. Measures of polygyny in humans. Current Anthropology. 29: 189-94.

47. Lu J, Wu CI. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci 102: 4063-7.
48. Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23: 23-35
49. McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Kidd JM, Wall JD, Bustamante CD et al. 2015. Inference of gorilla demographic and selective history from whole-genome sequence data. Mol Biol Evol 32: 600-12.
50. McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 8;5:e1000471.
51. Møller AP. Ejaculate quality, testes size and sperm competition in primates. Journal of Human Evolution. 1988 Aug 31;17(5):479-88.
52. Nam K, Munch K, Hobolth A, Dutheil JY, Veeramah KR, Woerner AE, Hammer MF, Mailund T, Schierup MH, Prado-Martinez J, et al. 2015. Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc Natl Acad Sci 112: 6413-8.
53. Nascimento JM, Shi LZ, Meyers S, Gagneux P, Loskutoff NM, Botvinick EL, Berns MW. 2008. The use of optical tweezers to study sperm competition and motility in primates. Journal of the Royal Society Interface. 5: 297-302.
54. Nicolaisen LE, Desai MM. 2012. Distortions in genealogies due to purifying selection. Mol Biol Evol 29: 3589-3600.
55. Nicolaisen LE, Desai MM. 2013. Distortions in genealogies due to purifying selection and recombination. Genetics. 195: 221-30
56. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.
57. Nguyen LP, Galtier N, Nabholz B. 2015. Gene expression, chromosome heterogeneity and the fast-X effect in mammals. Biol Lett. 11: 20150010.
58. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499-502.
59. Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution.61: 3001-6.
60. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. Nature 499: 471-5.
61. Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight JR. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature. 486: 527-31.
62. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-2.
63. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 328: 636-9.
64. Schrider DR, Kern AD. 2015. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. Genome Biol Evol 7: 3511-28.
65. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-50

66. Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF, Bustamante CD, Hammer MF, Wall JD. 2016. The time-scale of recombination rate evolution in great apes. Mol Biol Evol. 33: 928-45.
67. Surbeck M, Mundry R, Hohmann G. 2011. Mothers matter! Maternal support, dominance status and mating success in male bonobos (Pan paniscus). Proceedings of the Royal Society of London B: Biological Sciences. 278: 590-8.
68. Tachida, H. 2000. DNA evolution under weak selection. Gene 261: 3-9.
69. Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. 2006. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. Mol Biol Evol. 23:565–73
70. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, Sninsky JJ, Cargill M, Adams MD, Bustamante CD et al. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 5: e1000592.
71. Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. 2014. Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. Mol Biol Evol. 31:2267-82.
72. Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, McVean G. 2014. Strong male bias drives germline mutation in chimpanzees. Science. 344: 1272-5.
73. Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. Nat Genet. 27: 422-6.
74. Watts DP. 1988. Coalitionary mate guarding by male chimpanzees at Ngogo, Kibale National Park, Uganda. Behavioral Ecology and Sociobiology. 44: 43-55.
75. Watts DP. 1990. Mountain gorilla life histories, reproductive competition, and sociosexual behavior and some implications for captive husbandry. Zoo Biology. 9: 185-200.
76. Watts DP. 1991. Mountain gorilla reproduction and sexual behavior. American Journal of Primatology. 24: 211-25.
77. Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. Mol Biol Evol 19: 1376-84.
78. Wilson ML, Boesch C, Fruth B, Furuichi T, Gilby IC, Hashimoto C, Hobaiter CL, Hohmann G, Itoh N, Koops K, Lloyd JN. 2014. Lethal aggression in Pan is better explained by adaptive strategies than human impacts. Nature. 513: 414-7.
79. Wrangham RW. 1993 The evolution of sexuality in chimpanzees and bonobos. Human Nature. 4: 47-79.

APPENDIX C

FASTER METRIC NEAREST NEIGHBOR SEARCH USING DISPERSION TREES

Intended as a submission to the Proceedings of the VLDB Endowment

Abstract

We introduce a data structure called a dispersion tree to solve the k-nearest neighbor (k-NN) problem in quasi-metric and metric spaces. Dispersion trees are a hierarchical data structure based on a large dispersed set of points, and unlike previous works, they are created using both top-down and bottom-up construction techniques. Our bottom-up construction uses a nontrivial objective function that we compute exactly, with the objective being the formation of balls with small radii. We also introduce a 1/4 approximation algorithm for find a dispersed set in metric spaces, and we provide a method for top-down k-NN search that is performs the minimum number of distance computations needed to search any metric tree. Dispersion trees have faster query-times than comparable data structures in both structured and unstructured metric spaces in the vast majority of experimental conditions examined, all whilst maintaining a worst-case construction time that is sub-quadratic.

0.1 Introduction

Nearest-neighbor (NN) search is a generalization of Knuth's classic post office problem [14]. In Knuth's presentation each residence can query its closest (1-NN) post office. In the k-NN generalization of the problem, a query is assigned to its k closest points contained in some dataset. k-NN search is seen in a multitude of applications, ranging from the NN search in Euclidean space implemented in PostgreSQL, to the k-NN machine learning classification technique in a more generalized feature space. Motivating our work on the subject is the use of k-NN search in protein databases, where we use k-NN classification to predict protein secondary structure. The biological sciences provide obvious examples of extremely large databases that perform nearest-neighbor searches, with the national center for biological information's sequence read archive surpassing 5.6 PetaBases[1] as of June, 2016.

While the need for efficient k-NN search is seen in a variety of applications and disciplines, efficient data structures for performing exact k-NN search remains an area of active research. Optimizations beyond exhaustive search can be found when the problem is restricted to metric spaces, where in general the triangle inequality can be leveraged to efficiently organize the points with metric data structures (See Section 2). Metric data structures nearly universally perform top-down construction, where the points in the dataset are recursively partitioned in accordance to their distance to a constant number of reference points. While this approach is intuitive and perhaps appropriate in a low-dimensional setting, we argue that top-down partitioning techniques have the unfortunate property that some, and perhaps many, points that

[1] http://www.ncbi.nlm.nih.gov/Traces/sra/

are close in the metric space are placed far from each other in the index of the metric space. We posit that this stems from the weak requirement of the triangle inequality, where partitioning many points with respect to their distance to few references will always place points that are relatively "far" from all references in largely arbitrary partitions; an effect that is exacerbated with each recursive call. In high dimensional spaces the vast majority of points are in fact far apart, implying that this byproduct of the "curse of dimensionality" may also "curse" top-down construction techniques.

This paper presents a data structure called a dispersion tree that has a single, nonrecursive partitioning stage to a large ($\sqrt{n}$) number of reference points. The points within the blocks, and later the subtrees that span the blocks, are then greedily merged bottom up. Subtrees are merged under the criterion of minimum radius using a computation that does not rely on the triangle inequality, but is instead exact. Despite this exactness, the two-stage design of dispersion trees (Figure 2) insures that the construction times are subquadratic, making it an attractive search tool for datasets of moderate to large size.

0.1.1 Background

We begin by defining some basic terminology. A metric space is an ordered pair (M, d), where d is a distance function defined over all pairs of objects in set M. d is constrained $\forall x, y, z \in M$ to

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \iff x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

. Removal of the third rule results in a generalization of a metric space termed a quasi-metric space.

Metric spaces are seen in a variety of disciplines, ranging Euclidean, Manhattan and the more general Minkowski distance on vectors, to the Levenshtein and Hamming distances on strings, to name a few.

Metric data structures require an initial investment in computation time on the organization of points in a metric space into a structure that allows k-NN queries to be performed quickly. While quantifying speed in units of time is a deep and sophisticated topic that has machine-dependent properties, the speed of k-NN search in metric spaces is typically quantified in distance computations i.e., the number of times that the distance function was invoked. Pruning can be used to reduce distance computations, wherein the distance of the query to a single point in a subset of the dataset can be used to infer a lower-bound distance between that subset and the query. If this lower bound is sufficiently far, then it can be pruned. Pruning in metric

spaces is typically either based on a hyperplane or based on a ball. For the latter, a subset of points in a metric space have a center point (p), and all other points in that subset being at most distance R to p. For a query q, we can use the triangle inequality to arrive at a lower bound on the distance from q to any point in the ball. Specifically, this lower bound is the distance from p to q ($d(p,q)$) $-R$. Thus, if q's current k-NN distance (r) is less than this lower bound, the ball can be pruned and no further distance computations are required (Figure 1). Note that for a fixed q, r, and p, the sole factor that dictates whether or not pruning occurs is R. Further, if *a priori* all p are equally likely neighbors of q, then for the purposes of this argument they are effectively fixed, and if optimal search strategies are used (e.g., [13]), then r is minimized, fixing this quantity as well. Thus the sole outstanding property that dictates pruning is R, and it is R that dispersion trees, unlike any other metric space index, attempt to greedily minimize from the bottom up.

0.1.2 Related Work

k-NN search in an arbitrary metric spaces is a well-studied problem, with excellent reviews on the subject seen in [[20], [9]]. k-NN search can either be solved approximately or exactly. Approximate NN (ANN) search methods may project the original high-dimensional points into a lower-dimensional space, and then use exact k-NN search in this simpler search space [21], or they may employ approximate search strategies of data structures that may otherwise return exact results [[1], [8]]. Another popular ANN search approach is to use locality sensitive hashes (LSH) [10]. There are many variants of the LSH [[19] [16], [22], [12] to name a few], with more recent works focusing on reducing the LSH's otherwise costly IO requirements [15]. While effective within their own problem statements, approximate methods may give inexact results, and the ability to tolerate these errors is domain specific.

Exact k-NN search structures can be broken down into two basic subtypes. Matrix-based approaches, such as Vidal's approximating and eliminating search algorithm

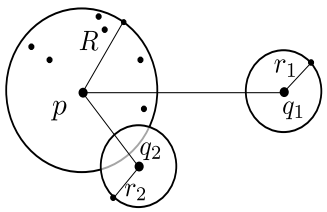


FIGURE 1. *Ball pruning.* The ball centered on point p with radius R can be pruned if query q is sufficiently far away. Specifically, as in the case of q_1, if the distance $d(p,q) \geq R + r$ then no point in p's ball can yield a closer nearest neighbor. The case of q_2 may yield a nearest neighbor as the two balls "overlap", and as such the subtree spanning that ball must be explored.

[24], and further developments by [17], produce k-NN search results in a constant number of evaluations of the distance function (i.e., distance computations). k-NN search nevertheless takes $\Omega(n)$ time, where n is the size of the database, implying that these approaches are best reserved for a distance function that is costly. Tree-based approaches were first introduced for discrete metrics in [5], and more generally by Uhlmann [23]. Uhlmann introduced what would later be called vantage point (vp) trees [25], as well as generalized hyperplane trees. With vp-trees a single reference, or vantage point, is chosen, and the median-sphere centered at this vantage point is used to create a bipartition. The subset inside and outside of the median-sphere are recursed upon until single points remain. Extensions of vp-trees are seen with mvp-trees [3], which partitions according to not just one but multiple vantage points, and allows the partitioning to occur at multiple order statistics as opposed to just medians. Generalized hyperplane trees use not one but two pivots, with the partitioning being formed by assigning the remaining points to whichever pivot they are closer too, forming a "hyperplane", with both sides of the hyperplane being recursed upon. The GNAT design [4] extends this idea to an arbitrary arity a, with precomputing $O(a^2)$ bounds in each node to aid in search. Spatial approximation trees [18] are a type of GNAT, where the differences between them lie in the construction. Specifically, in a GNAT the pivots are chosen to be mutually far from each other, while with the spatial approximation tree the pivots are chosen to be mutually closer to some center point than to any other pivot. Antipole trees provide a blend of generalized hyperplane trees and vp-trees. With antipole trees, the dataset is hyperplane partitioned until balls of sufficiently small radius can be formed [6].

While the preceding data structures recursively partition points based on their distance to a few (typically 1 or 2) reference points, the list of clusters [7] is a simple yet surprisingly effective data structure that uses a single-pass partitioning strategy. Specifically, the list of clusters iteratively removes subsets (balls) with a specified number of points closest to a given reference point. The balls are organized in a list; if a particular ball cannot be pruned then it is exhaustively searched, and k-NN search continues until either the list is pruned or exhausted. Of note, the list of clusters is the only metric search data structure (that we know of) that does not use divide and conquer recursion to subdivide the dataset, and despite its simplicity it is often a competitive data structure in high dimension.

Cover trees [2] provide an interesting theoretical perspective on nearest neighbor search in metric spaces. Cover trees are level trees where the radius (R) of balls is reduced by a constant fraction (set to 1.3 in practice), and the points seen within any given level are required to be mutually far apart. These constraints allow for some strong asymptotic guarantees on both tree construction and NN search. For metric datasets of size n with an expansion constant c, 1NN search time is $O(c^{12} \log n)$, while construction time is $O(c^6 n \log n)$ and space is $O(n)$. For points in a uniform d-dimensional vector space $c \sim 2^d$, which means that while the theoretical contribu-

tions of cover trees are innovative, their search-times in a practical setting may be dominated by these constants.

0.1.3 The Dispersion Tree

Our first goal in the design of the dispersion tree was to reconsider what properties a metric space index should have. The first property that we propose is that points that are close in the metric space should also be close in the index of the metric space. We posit that top-down approaches, where large numbers of points are partitioned with respect to their distance to a constant number of reference centers, will tend to place points that are relatively "far" from all references in largely arbitrary partitions. As points in high dimensional spaces are in general far apart, this implies that in high-dimensional settings, top-down approaches will tend to scatter points that are close in the metric space throughout the index of the metric space. Bottom-up approaches, on the other hand, have the ability to place nearest neighbors together, begetting the formation of balls with small radii (i.e., R in Figure 1). The second design decision of note is the use of a non-trivial objective function on the merging of subtrees. We further argue that the objective function should be an exact computation, and not one stemming from bounds from the triangle inequality, as these bounds are in general quite loose. As the number of possible mergers for a set of n subtrees is $O(n^2)$, evaluating $\sqrt{n}$ subtrees at a time is a natural fit to the formulation.

With these principles in mind, we present the dispersion tree. The algorithms used

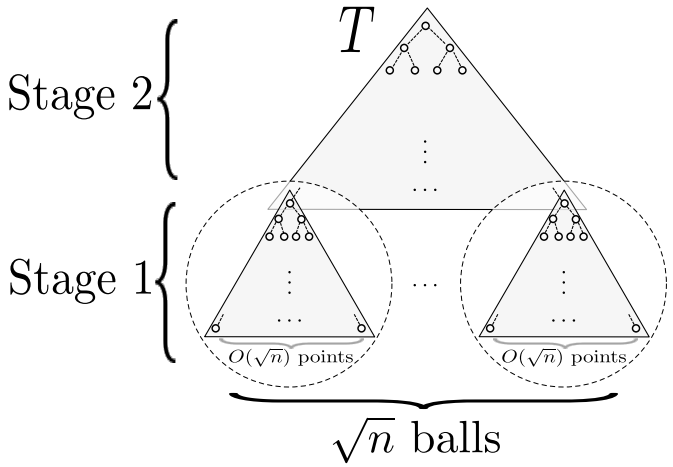


FIGURE 2. *The anatomy of a dispersion tree.* Dispersion trees (T) are ball trees that have two distinct construction phases. In stage 1, after partitioning the points into $\sqrt{n}$ blocks, the points are treated as leaves and the subtrees in each block are greedily merged under the objective of minimum radius until a single subtree remains. In stage 2 the balls that span each block are greedily merged using the same criterion.

to create a dispersion tree are seen in the next section (beginning with `CreateDispersionTree`). Dispersion trees are constructed over a dataset (D), and begin with the creation of a `Partition` based on $\sqrt{|D|}$ references.

Selecting reference points requires some care, and several stages of processing. First, we form an `ApproximateDispersedSet` centers (C) of size $\sqrt{|D|}$. The approximate dispersed set algorithm greedily removes arbitrary points, and the points in $|D|$ they are farthest from, until sufficient numbers of centers are found. This dispersed set is then polished by considering the remaining points in $D - C$, greedily evaluating them to maximize the average (sum) of the distances between the points in C. We then form a hyperplane partition on D, assign all points to their closest center. This leaves us with two possible problems; the sizes of the blocks formed by hyperplane partitioning are arbitrary, while our time analysis requires the sizes of the blocks to be of size $O(\sqrt{|D|})$. Second, the centers in C may not be "central". We address the latter by using our `Center` routine to find a `MinSum` center for each block of the hyperplane partition. To address the former we then sort each point by its minimum distance to any of the centers, and under this ordering we assign points to the block associated to its closest center, conditioning on that center's block not have more than $c \cdot \sqrt{n}$ points in it, setting c to 2 in practice.

After we form our partition (P), the points within each block are initialized to be leaves, i.e., balls with radius 0 and have no children. The leaves within each block of the partition are then greedily merged with the `Merge` routine into a single tree that spans that block. In a dispersion tree, merging subtrees i and j is an asymmetric operation wherein the jth subtree is merged into the ith, with the ith subtree's center being the center of the new parental tree. The radius of this subtree is simply the $\max(i.radius, \max_{\forall(l \in points\ in\ j)}(distance(l, i.center)))$.

Merging begins by first considering all possible radii of all possible mergers of leaves. Specifically, $R[i, j]$ contains the radius of the resulting subtree if subtree j is merged into subtree i. In the case of leaves, this is equivalent to the distance between the centers of i and j. When the jth subtree is merged into the ith, subtree j is no longer considered to be current, and $Current[j]$ is set to $false$. The contents of R are loaded into a min-heap H, which allows us to extract which subtrees, when merged, will yield the smallest radius, with the first such merger being the two closest nearest neighbors in the block. Once the merger is chosen, the bounds in $R[i, j]$ are updated to reflect the new radii. Note that this update only involves the columns of R, that is, the values in there table where the new subtree i's radius increased, to reflect future possible mergers of i into some other subtree j. This update is not necessary for the rows of R, that is the balls where the new subtree i's center will be maintained, as we are taking a max when we update R, and the value from H is the minimal. Rather than update the heap H immediately, which would take an augmentation to H, we instead choose a lazy update scheme. The radii in $R[i, j]$ are monotonic increasing and are always correct. The values in H may become stale (i.e., $< R[i, j]$), thus H is

lazily updated when stale values are pulled from the heap, and only if the merger is still valid (neither subtree has been consumed). This process continues until there is exactly one subtree left, which `Merge` returns.

The final subtrees that span each block of the partition are then run through the same `Merge` procedure, with the final tree being the root for the entire dataset. We found that three practical modifications to our routines yield modest improvements in runtime. The first is that in the case of leaves with the `Merge` procedure we use $R[i, j]$ to find the center that minimizes the radius of the entire block of the partition, and we ensure that that center is the center used in the final tree we return. Second, as many distance functions are symmetric (noting that dispersion trees do not require symmetry), we use a tie-breaking rule in H, wherein the case of ties centers with the smaller average distance to the other points being considered are chosen. Third, we find that using centers from the approximate dispersed set, compared to arbitrary points, actually increases our search times. This may be because these points, while appropriate for the maximum dispersion problem, make poor centers for forming partitions. As such, we use arbitrary centers as initial values, prior to partitioning and applying our center-finding heuristics.

To perform k-NN search we use the `NearestNeighborSearch` algorithm. This is variation of the "range-optimal" incremental search algorithm of Hjaltason and Samet [13], which searches no more of the dispersion tree than would a range-search started with the range of the kth nearest neighbor distance.

0.2 Algorithm Descriptions

procedure `CreateDispersionTree`(D) **begin**
    $P :=$ `Partition`$(D, \sqrt{|D|}, c * \sqrt{|D|})$
    for every block B_i in partition P **do begin**
        Let L_i be a set of leaves corresponding to the points
            in block B_i
        $T_i :=$ `Merge`(L_i)
    end
    $T :=$ `Merge`$(\left\{T_1, T_2, \ldots, T_{\sqrt{|D|}}\right\})$
    return T
end

procedure `Merge`(T) **begin**
    Create an empty Min Heap H
    Create array $R[1:|T|, 1:|T|]$ that will hold bounding-
        radii $R[i, j]$ for merging subtrees T_j into T_i in T
    Create a boolean array Current$[1:|T|]$ with
        all entries initialized to *true*

```
    Create array Node[1:|T|] with entries initialized to the
        roots of the trees in T
    for all distinct i, j ≤ |T| do begin
        Let P_j be the set of points at the leaves of T_j
        R[i, j] := max(Node[i].radius, radius(Node[i].center, P_j))
        Insert((R[i, j], i, j), H)
    end
    m := 0
    while m < |T| − 1 do begin
        (r, i, j) := ExtractMin(H)
        if Current[i] and Current[j] then begin
            if r < R[i, j] then
                Insert((R[i, j], i, j), H)
            else begin
                Initialized a node v to have radius r, the same
                    center as Node[i] and children set to Node[i]
                    and Node[j]
                Node[i] := v
                Current[j] := false
                m := m + 1
                for k := 1 to |T| other than i
                    if Current[k] then
                        R[k, i] := max(R[k, i], R[k, j])
                end
            end
        end
    end
    Create tree t whose root is the single current node
        in Node
    return t
end

procedure Partition(S, m, l) begin
    C := ApproximateDispersedSet(S, m)
    for every p in S − C do begin
        Find the point c in C for which swapping p with c
            gives the greatest increase in the sum of all
            pairwise distances.
        Swap p with c in C if swapping p with c increases
            this sum
    end
    Form a hyperplane partition P by assigning all of S to
```

```
        their closest center in C
    C := {}
    Let a and b be appropriate constants for the center
        finding heuristic.
    for every block B in P do
        C := C ∪ Center(B, a, b)
    sort the points in S by their minimum distance to a
        center in C
    for each point p in S in sorted order do
        Assign p to the closest center's block that does not
            have l points in it
    return this assignment
end

procedure ApproximateDispersedSet(S, k) begin
    D := {}
    for i := 1 to ⌊k/2⌋ do begin
        Pick an arbitrary point v from S
        Find the farthest point w from v in S
        D := D ∪ {v, w}
        S := S − {v, w}
    end
    if k is odd then begin
        Find the point v in S that maximizes the sum of the
            distances to the points in D
        D := D ∪ {v}
    end
    return D
end

procedure Center(D, r, s) begin
    if |D| ≤ s then
        return MinSum(D, D)
    Form partition P by randomly partitioning D into r
        equal-sized groups
    C := {}
    for every block B in partition P do
        C := C ∪ {Center(B, r, s)}
    return MinSum(C, D)
end

procedure NearestNeighborsSearch(k, q, T) begin
```

```
    Initialize a min-heap H, and a max-heap K
    d := D(q, T.Root.center)
    Insert((0, d, T.Root), H)
    Insert((d, T.Root.center), K)
    while H is not empty do begin
        (b, d, v) := ExtractMin(H)
        if b ≥ Maximum(K) and Size(K) = k then
             break
        if Size(K) < k or
            b < Maximum(K)  then begin
            for each child w of v do begin
                if w.center ≠ v.center then begin
                    e := D(q,w.center)
                    if Size(K) < k then
                        Insert((e,w.center), K)
                    else if e < Maximum(K) then begin
                        ExtractMax(K)
                        Insert((e,w.center), K)
                    end
                end else
                    e := d

                if Size(K) < k  or
                    Maximum(e−w.radius, b) < Maximum(K) then
                    Insert((Maximum(b, e−w.radius), e, w), H)

            end
        end
    end
    return the points in K
end
```

0.2.1 Time Analysis

Theorem 1. *`Merge(t)` takes $O(|t|^2 \log |t|^2 + |t|n)$ time for a set of trees t composed of n total leaves*

Proof. The array of bounding radii $R[i, j]$ and the max-heap H are initialized to the maximum distance from the center of every tree in t to the leaves in every remaining tree in t, which takes $O(|t|n)$ time. Then, $|t| - 1$ subtrees are merged, and $O(|t|)$ radii in $R[i, j]$ are updated after each merge, which takes $O(|t|^2)$ time in total. The heap values are lazily updated in the **if**, and the number of lazy updates is bounded by the

number of updates in the **else** to $R[i, j]$. As each update to H takes $O(\log |t|^2)$ time, this yields a running time of $O(|t|^2 \log |t|^2)$ for the **for** loop. Finding the last current tree in t takes $O(|t|)$ time. In total, `Merge` takes
$O(|t|^2 \log |t|^2 + |t|n)$ time. □

Theorem 2. *Constructing a dispersion tree takes $O(n^{1.5} \log n)$ time and $O(n)$ space for a database of n points.*

Proof. Let m be $\sqrt{n}$. `Partition` begins by making an approximate dispersed set of size m, which takes $O(n^{1.5})$ time. The dispersed set is polished by greedily considering a swap with every remaining point in the database, which with the maintenance of key subsums takes $O(n^{1.5})$ time. To form the hyperplane partition P, all $n-m$ points are compared to m centers, and then partitioned using a counting sort, which takes $O(n^{1.5})$ time. Taking r and s as constants, by the Master Theorem `Center` takes $O(l \log l)$ time for any dataset of size l. Let n_i be the size of the ith block in P. It follows that

$$\sum_{i=1}^{m} n_i = n$$

. As such, the time needed to call `Center` on the blocks in P is:

$$\sum_{i=1}^{m} n_i \log n_i \ \leq \sum_{i=1}^{m} n_i \log n \ = n \log n$$

The final partition is based off of computing the distance of each of the n points to the m centers, sorting according to this distance, and then greedily partitioning based off of the ordering of the sort, which is $O(n^{1.5}) + O(n \log n) + O(n^{1.5})$.

By *Theorem 1*, running `Merge` on each block takes $O(m^2 \log m^2 + m^2)$ time, which is $O(n \log n)$, and as there are m blocks, this takes $O(n^{1.5} \log n)$ total time. `Merge` is run one final time on the m subtrees, which by *Theorem 1* is $O(m^2 \log m^2 + mn)$, which is $O(n^{1.5})$. In total, it follows that it takes $O(n^{1.5} \log n)$ time to create a dispersion tree.

Dispersion trees themselves take $O(n)$ space, as each point is present at the leaves, and each internal node uses exactly 1 of the 2 child centers as its center, giving $2n-1$ nodes in total. As each node uses a constant amount of space, a dispersion tree takes $O(n)$ space. It is plain that `Partition` takes $O(n)$ space. `Merge` on a set of trees t takes $O(|t|^2)$ space for the array $R[i, j]$ and heap H, however it is only ever called on $\sqrt{n}$ subtrees at a time, which takes $O(n)$ space. Thus, it follows that it takes $O(n)$ space to create a dispersion tree. □

0.3 Approximation algorithm for Maximum Dispersion

In our context, the *Maximum Dispersion Problem* is, given a metric space (S, d) and an integer $k \geq 0$, find a subset $D \subseteq S$ of cardinality $|D| = k$ that maximizes the sum of the distances $\sum_{p,q \in D} d(p, q)$.

While Maximum Dispersion is NP-complete, so efficiently finding an optimal set is unlikely, we can efficient find a *near-optimal* set. An *α-approximation algorithm* for a maximization problem, where $\alpha < 1$, is a polynomial-time algorithm that is guaranteed to find a solution whose value is at least factor α times the optimum.

Recall that our heuristic for Maximum Dispersion starts with the empty set D, and repeatedly performs the following $\lfloor k/2 \rfloor$ times: pick an arbitrary point $p \in S - D$, find a farthest point $q = \text{argmax}_{r \in S-D} d(p, r)$, and add $\{p, q\}$ to D. Finally if k is odd, add to D an arbitrary point from $S - D$.

Theorem 3 (Approximating Maximum Dispersion)**.** *Procedure* `ApproximateDispersedSet` *is an α-approximation algorithm for Maximum Dispersion with*

$$\alpha \;=\; \frac{1}{4}\left(1 \,+\, \frac{1}{2k-3}\right).$$

Proof. For a given integer i, let

- D_i^* be an optimal dispersed set of cardinality i,
- $\widetilde{D}_i$ be the dispersed set of cardinality i,
- $\overline{e} = (p, q)$ be the farthest pair of points in S,
- $\widetilde{e} = (v, w)$ be the first pair of points v, w added to D by the heuristic run on S,
- S' be $S - \{v, w\}$,
- $\widetilde{D}_{i-2}$ be the dispersed set of cardinality $i - 2$ found by the heuristic on S',
- D_{i-2}^* be an optimal dispersed set of cardinality $i - 2$ on S',

and for a set D, let $f(D)$ be the objective function for Maximum Dispersion, namely $\sum_{p,q \in D} d(p, q)$.

We first show that $d(\widetilde{e}) \geq \frac{1}{2} d(\overline{e})$. Notice that $d(\overline{e}) = d(p, q) \leq d(p, v) + d(v, q)$ by triangle inequality, where $\widetilde{e} = (v, w)$. Furthemore, $d(p, v) + d(v, q) \leq 2d(v, w)$, as w is farthest from v. Combining these two inequalities gives $d(\overline{e}) \leq 2d(\widetilde{e})$. Thus at each iteration, our heuristic finds an approximate farthest pair in $O(n)$ time.

We map the pair $\widetilde{e} = \{v, w\} \subseteq \widetilde{D}_i$ to a pair $e^* = \{x, y\} \subseteq D_i^*$ as follows.

- If $|\{v, w\} \cap D_i^*| = 2$, then $e^* = \widetilde{e}$.

- If $|\{v, w\} \cap D_i^*| = 1$, then we take for x the element that is in common between $\{v, w\}$ and D_i^*, and for y we take an arbitrary element of $D_i^* - \{x\}$.
- If $|\{v, w\} \cap D_i^*| = 0$, then for x, y we take an arbitrary pair of elements in D_i^*.

Note that with this mapping of v, w to x, y, we have $D_i^* - \{x, y\} \subseteq S - \{v, w\} = S'$.

We now prove $f(\widetilde{D}_k) \geq \alpha f(D_k^*)$ by induction on k. Our basis is $k = 1, 2$. For $k = 1$, notice that $f(\widetilde{D}_1) \geq \alpha f(D_1^*)$, as for a set of cardinality 1 the objective f is 0. For $k = 2$, notice that $f(\widetilde{D}_2) \geq \frac{1}{2} f(D_2^*)$, by our earlier bound $d(\widetilde{e}) \geq \frac{1}{2} d(\overline{e})$, and $\alpha = \frac{1}{2}$ for $k = 2$. So the basis holds.

For the inductive step with $k > 2$, assume the theorem holds for all $k' < k$. Now,

$$
\begin{aligned}
f(D_k^*) &= d(e^*) + \sum_{z \in D_k^* - \{x,y\}} \big(d(x,z) + d(y,z)\big) \\
&\quad + \sum_{\text{distinct } z,z' \in D_k^* - \{x,y\}} d(z,z') && (1) \\
&\leq d(\overline{e}) + 2(k-2) d(\overline{e}) + f(D_{k-2}^*) && (2) \\
&\leq 2\big(1 + 2(k-2)\big) d(\widetilde{e}) + \frac{1}{\alpha} f(\widetilde{D}_{k-2}) && (3) \\
&= \frac{1}{\alpha}\big(1 + (k-2)\big) d(\widetilde{e}) + \frac{1}{\alpha} f(\widetilde{D}_{k-2}) && (4) \\
&\leq \frac{1}{\alpha}\Bigg(d(\widetilde{e}) + \sum_{z \in \widetilde{D}_{k-2}} \big(d(v,z) + d(w,z)\big) \\
&\quad + f(\widetilde{D}_{k-2}) \Bigg) && (5) \\
&= \frac{1}{\alpha} f(\widetilde{D}_k)\,. && (6)
\end{aligned}
$$

In the above, equation (1) follows from the definition of f. Inequality (2) follows from the fact that $\overline{e}$ is a farthest pair in S, and $f(D_k^* - \{x, y\}) \leq f(D_{k-2}^*)$ since $D_k^* - \{x, y\} \subseteq S'$. Inequality (3) follows from $d(\overline{e}) \leq 2d(\widetilde{e})$, and by our induction hypothesis on k, as $D^*k - 2$ and $\widetilde{D}_{k-2}$ are both over set $\widetilde{S}$. Equation (4) follows from the fact that $\frac{1}{\alpha}\big(1 + (k-2)\big) = 2(2k-3)$ for our particular approximation ratio α. Inequality (5) follows from the triangle inequality, and equation (6) follows from the definition of f.

Thus approximation ratio α holds. By our earlier analysis, procedure `ApproximateDispersedSet` runs in polynomial time, so it is an α-approximation algorithm. □

0.4 Optimally searching the tree

We call the algorithm used by `NearestNeighborSearch` on the dispersion tree, *least-lower-bound search*, as it chooses for the next node of the search tree T to visit the one

of least lower-bound value. It does not appear to be widely known that this search algorithm is essentially *optimal*, in the sense of using the fewest possible number of distance computations to find the k-nearest neighbors of a query point with a given tree T. The various approaches in the literature for k-nearest-neighbor search employ a variety of strategies for searching the trees they construct (such as variants of depth-first and breadth-first search). In this section we prove that least-lower-bound search, modulo ties in lower bounds, is optimal.

For the optimality result, we need to correctly handle ties in lower bounds at tree nodes. The version of least-lower-bound search that we prove is optimal uses a *tie-breaking oracle*: faced with several equally-good nodes to choose among that all have the same lower-bound value, the oracle selects the tied node that results in the fewest total distance computations. A tie-breaking oracle is necessary for optimality: in the worst case, the tree could provide the degenerate lower-bound 0 at every tree node, forcing least-lower-bound search to blindly traverse the tree. In practice, for real-valued distance functions such as the metric space $\mathcal{R}^d$ under Euclidean distance, ties are rare. When all tree nodes have distinct lower-bound values, our result shows that ordinary least-lower-bound search—without an oracle—is in fact optimal.

We prove optimality among the class of *top-down* search algorithms: when the point at a tree node is revealed to such an algorithm, it must have also visited all the ancestors of the node. This restriction is necessary as well: without it, when the query point q is itself in T, an omniscient algorithm could jump to the node containing q, achieve a query radius of $d(q, q) = 0$, and then prune the root of T—with which least-lower-bound search cannot compete.

We also require the search algorithms all use the same *evaluation rule* for distance computations along paths of tree nodes. For example, procedure `NearestNeighborSearch` recognizes when a parent and child share the same center point, to save a distance computation at the child. Any such rule for saving distance computations along paths is used by all search algorithms in the class.

The effect of these restrictions is that the key difference among search algorithms is the order in which they visit nodes in T, which affects the value of their query radius when attempting to prune a node.

Theorem 4 (Optimal tree search)**.** *Consider any search tree T, any evaluation rule for saving distance computations along paths in T, and all top-down algorithms for k-nearest-neighbor search using this rule on T. With respect to the total number of distance computations, least-lower-bound search with a tie-breaking oracle is optimal.*

Proof. Let A^* be least-lower-bound search with a tie-breaking oracle, and A' be any other top-down algorithm for k-NN search. We prove that the total number of distance computations for A^* on T is at most the number for A' on T.

Consider the *iterations* of A^*, where each iteration removes the node from the heap wontith the smallest lower bound. We say the node removed at an iteration is

visited; this node is either *pruned* (due to the pruning inequality holding), or *explored* (which puts its two children on the heap).

Let v be the node removed at the *first* iteration where A^* explores v while A' does not (either because A' directly prunes v, or v is in a substree pruned by A'). Notice that for all prior iterations of A^*, either A^* prunes the node removed, or A^* and A' both explore the node. So if an iteration does *not* exist where A^* explores a node but A' does not, then the theorem holds (since for every node of T either they both explore it or A^* prunes it, so the total number of distance computations for A^* is at most that for A').

For the above node v, both A^* and A' explore the parent of v (else v is not the first such node). Thus A' also visits v (since it is top-down).

Since A^* explores v, we know $L(v) < R^*$, where $L(v)$ is the lower bound on the subtree for v and R^* is the query radius for A^* when it visits v. Since A' prunes v, we know $L(v) \geq R'$, where R' is the query radius for A' when it visits v.

Some of the points that lead to the smaller query radius R' for A' must come from the rest of T not yet processed by A^* (as if all these points come from the portion of T already processed by A^*, then $R' \geq R^* > L(v)$, contradicting $R' \leq L(v)$). Every point in the rest of T is in a subtree rooted at a node on the frontier for A^*, and every such frontier node has a lower bound at least as great as $L(v)$. Thus the points lacked by A^* that lead to R' are all at distance at least $L(v)$. This implies $R' \geq L(v)$. Recall that $R' \leq L(v)$, as A' prunes v. Hence $R' = L(v)$, and moreover, all these remaining points are at distance $L(v)$.

Each such point in the rest of T has an ancestor w on the frontier for A^*. Since v has minimum lower bound on the frontier, $L(w) \geq L(v)$, and since A' uses a point in the subtree of w for its query radius, $L(w) \leq R' \leq L(v)$. Together these imply $L(w) = L(v)$.

Thus the frontier for A^* has other nodes w that are tied with v, for which each such w has a path to a point of distance $L(v)$, a subset of which when merged with the current set for A^* will give an optimal k-NN set. As the lower bounds used by A^* are monotonic increasing on paths descending in T, all lower bounds along these paths equal $L(v)$, and no node off these paths has lower bound less than $L(v)$. Since A^* uses a tie-breaking oracle that minimizes the total number of distance function evaluations, in the remainder of its computation A^* will simply: (a) walk these paths of lower bound value $L(v)$, (b) collect these points of distance $L(v)$ to form its k-NN set, and then subsequently (c) prune away every remaining node of T.

Since A' is top-down, it must also walk these paths. Thus when A^* processes the remainder of T following the iteration for v, its number of distance computations is at most that for A'. By our earlier argument, this also holds for the portion of T processed prior to this iteration. Thus the total number of distance computations for A^* is at most that for A', which proves the theorem. □

0.5 Results

We contrasted the search performance of dispersion trees against a large sample of NN search trees in both structured and unstructured metric spaces. In all experiments, 1000 1-NN queries were performed, and the mean search time (in distance computations) is reported. Unless otherwise stated, all experiments vary in size from 2^{10} to 2^{20} by powers of 2.

0.5.1 Datasets

1. **Eucl** consists of points drawn from a unit hypercube under the Euclidean distance function.

2. **Prot** consists of unique kmers drawn NCBI's non-redundant protein database[2] using the `Blosum62` distance matrix.

3. **DNA** consists of unique kmers drawn from chromosome 1 of the hg19 human genome assembly[3] under the unit edit distance. Specifically, for each dataset size s with kmer size k, we use the first s overlapping unique kmers generated by using a windows size of k, sliding by one base at a time. For experiments where multiple independent runs were evaluated, chromosomes 1 through 10 were used.

0.5.2 Data Structures

A majority of the source code for comparable data structures come from the Similarity Search and Applications (SISAP) metric space library [11], from which we compared ourselves to generalized hyperplane trees [23], multivantage point trees [3], spatial approximation trees [18] and the list of clusters [7]. Source code for antipole trees [6] was obtained from the authors, and source code for cover trees [2] was found on the author's website[4]. Generalized hyperplane trees and multivantage point trees both require parameter tuning on the arity of the subtrees. To accommodate this we show the fastest results given an arity of 2,3,4,8 or 16. For the list of clusters and antipole trees we set their parameters to those suggested by the authors in [7] and [6], respectively. For the latter, the authors recommend using a cluster diameter 10% less than the median interobject distance value, which we estimated by sampling 1000 distances within each experimental condition. In cover trees, the radius of the balls shrinks by a constant fraction (we used the default of 1.3) between levels of the tree. Further, cover trees support batch as well as single queries, of which we report the latter as it had marginally better search times.

[2] ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz

[3] hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz

[4] hunch.net/ jl/projects/cover_tree/cover_tree.tar.gz

0.5.3 Timing Results

0.5.4 Unstructured Metric Spaces

We first sought to evaluate the relative performance of dispersion trees in an unstructured metric space. Using the **Eucl** dataset, we ran experiments on unit vectors of dimension 5, 10, 15 and 20 and computed the mean number of distance computations needed to solve 1-NN queries in the data structures described above (Figure 3). While dispersion trees are not the fastest in low dimension, in intermediate dimension they outperform all other data structures. Curiously, cover trees and dispersion trees both show better scaling in higher dimension rather than lower dimension. For dispersion trees, this may in part be a byproduct of the $O(\sqrt{n})$ height of dispersion trees, a worst-case behavior that may be more likely in lower dimensional settings. Or perhaps the greedy bottom-up ball merging strategy used by dispersion trees may create balls that are densely packed (i.e., more points per unit volume) when there are few points, but balls that are more loosely packed when there are more points owing to the reduction in the number of possible ways that balls with many points can be made. This in turn can create a situation where dispersion trees are less likely to prune balls that have many points in them, which is more likely to occur in lower dimensional settings.

0.5.5 Structured Metric Spaces

We next sought to evaluate the relative performance of dispersion trees in structured metric spaces. *A priori*, we anticipate that dispersion trees, as well as other data structures that try to exploit structure (e.g., spatial approximation trees), to exhibit fast search times in these benchmarks. Somewhat consistent with this expectation, dispersion trees exhibit the fastest search times across the structured benchmarks (Figures 4 and 5). As seen in Figure 4, not only are dispersion trees the fastest data structure, their search times are relatively invariant to the kmer size (i.e., the dimension). The design of the **DNA** experiment is meant to reflect the read mapping problem in the biological sciences, where short substrings (reads, which contain errors) are aligned to their closest matching position in a much larger string (the genome). In the **DNA** experiment, the 1-NN distance is at most 2, meaning that the 1-NN distance is never large, independent of the dimension of the space. This feature appears to be exploited by dispersion trees, and to a lesser extent by spatial approximation trees, while other data structures such as cover trees exhibit increasing search times with kmer size. In the **Prot** experiment, dispersion trees and cover trees are the fastest data structures, with cover trees and dispersion trees being relatively equally matched when the kmer size is large (e.g. $\mathbf{Prot}_{15}$) and with dispersion trees being faster when the kmer size is small (e.g. $\mathbf{Prot}_6$, Figure 5). Note that in the cases where cover trees and dispersion trees are equally matched, roughly 50% of the dataset is being searched, a case where an exhaustive brute-force search is probably the fastest

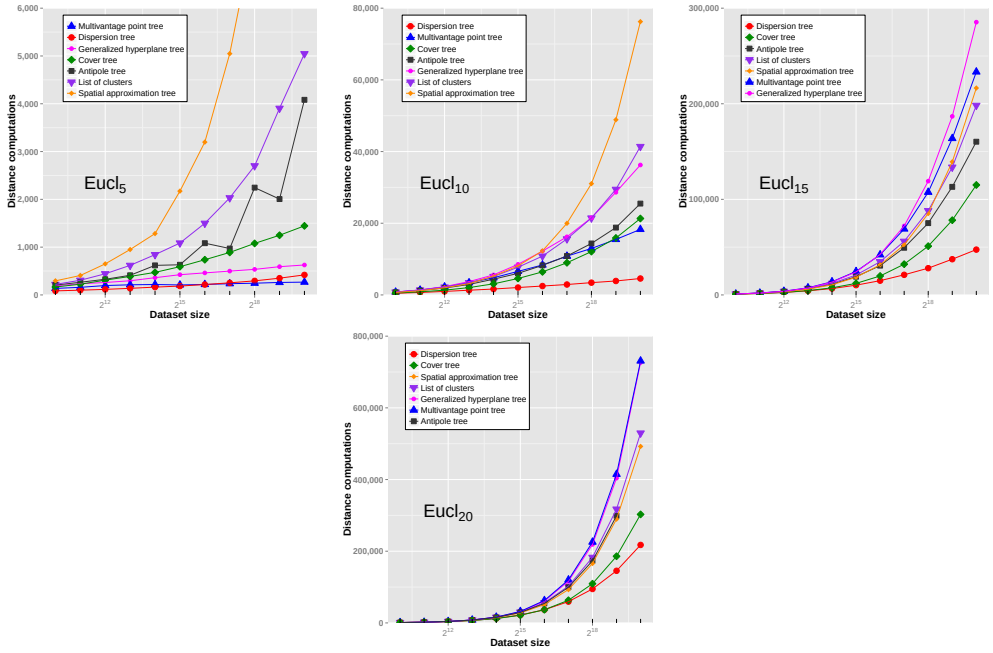


FIGURE 3. 1-*NN search times as a function of dataset size in unstructured metric spaces.* Using points taken from the `Eucl` dataset of dimension varying from 5, 10, 15 and 20 (left to right) dimensional space, 1-NN query times are contrasted against the size of the dataset (log scale). Note that the scale of the time axis varies between plots.

search-style in practice.

0.5.6 Scaling

While some of our timing results suggest that dispersion trees may exhibit polylog search times (e.g., $\mathbf{Prot}_6$ of Figure 5), whether or not our search times better fit a different order of growth remain to be seen. To empirically address this question we used nonlinear least squares regression using the `nls` function in the `R` programming language to fit our search times as a function of dataset size for the **Eucl** dataset (Figure 6) and the **Prot** dataset (Figure 7). We tested datasets ranging in size from 2^{10} to 2^{23}, by powers of 2, using 1000 queries for two data structures, dispersion trees and cover trees, as these data structures were generally fastest across our experimental conditions. We tested two functional forms: $y = a \cdot x^b$ and $y = a \cdot \log x^b$, and show the coefficient of determination (r^2), a measure of model fit, for both forms for both cover trees and dispersion trees, as well as the mean and standard deviation of the search times. Dispersion trees show marginally better fit for the $\log n^b$ search times across all but one experimental condition, however the level of support for n^b is also competitively high, especially as the kmer size / dimension grows. Cover trees show

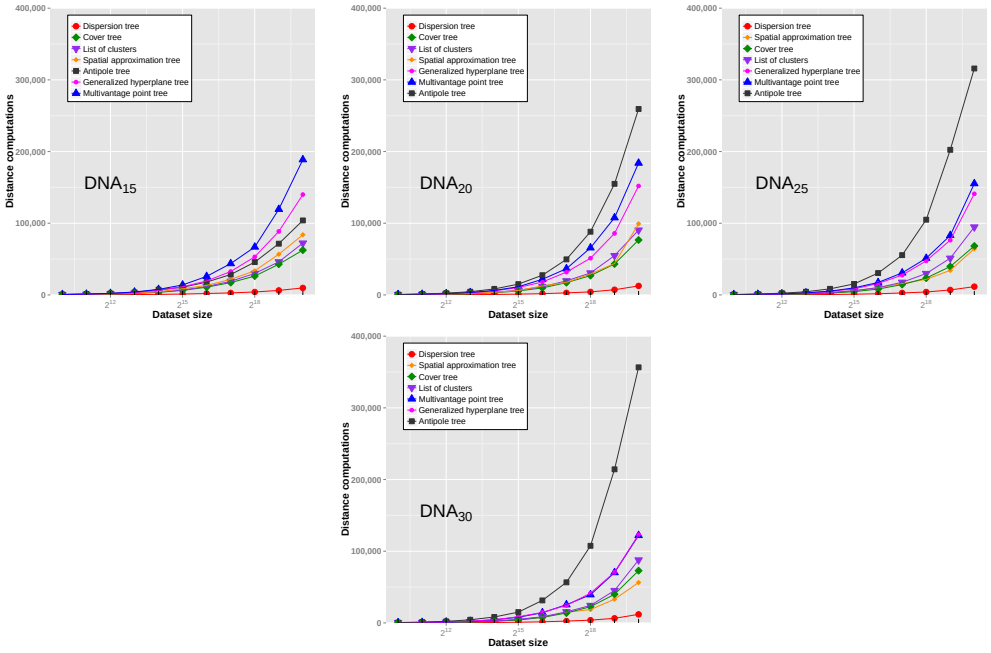


FIGURE 4. *The effects of dataset size on* 1*-NN search times in a structured metric space.* Mean 1-NN query times on the `DNA` dataset are shown as a function of the dataset size (x-axis) and kmer length (from left to right, 15 to 30). Differences from the left to right show the effects of kmer size (15 to 30). The plots are on a log-linear scale with a fixed y-axis.

.

higher levels of model fit in general, which may follow from the smaller variance in search times for cover trees. Dispersion trees fit a lower power (b) than cover trees, across functional forms and dimension / kmer sizes, which suggests that the faster search times in practice may not because of differences in lower order terms. The sole exception to these trends is the performance of dispersion trees in $\mathbf{Eucl}_5$ (Figure 6), which exhibits both better model fit for n^b than $\log n^b$, higher powers (b) for dispersion trees than cover trees, and perhaps most bizarrely, even higher powers for increased dimension ($\mathbf{Eucl}_5$ versus $\mathbf{Eucl}_{10}$, (Figure 6)). This phenomenon is unique to the **Eucl** dataset, and is not seen in **Prot** (Figure 7), suggesting that it is not low dimension *per se*, but unstructured low dimensional space that leads to the under-performance of dispersion trees. As the **Eucl** benchmarks are highly synthetic, this casts doubt on how meaningful this apparent lapse in performance actually is, especially as the absolute number of distance computations is still quite small even in the largest test case (a mean of 1081.6 versus 2192.3 distance computations for 2^{23} points in $\mathbf{Eucl}_5$ for dispersion trees and cover trees, respectively).

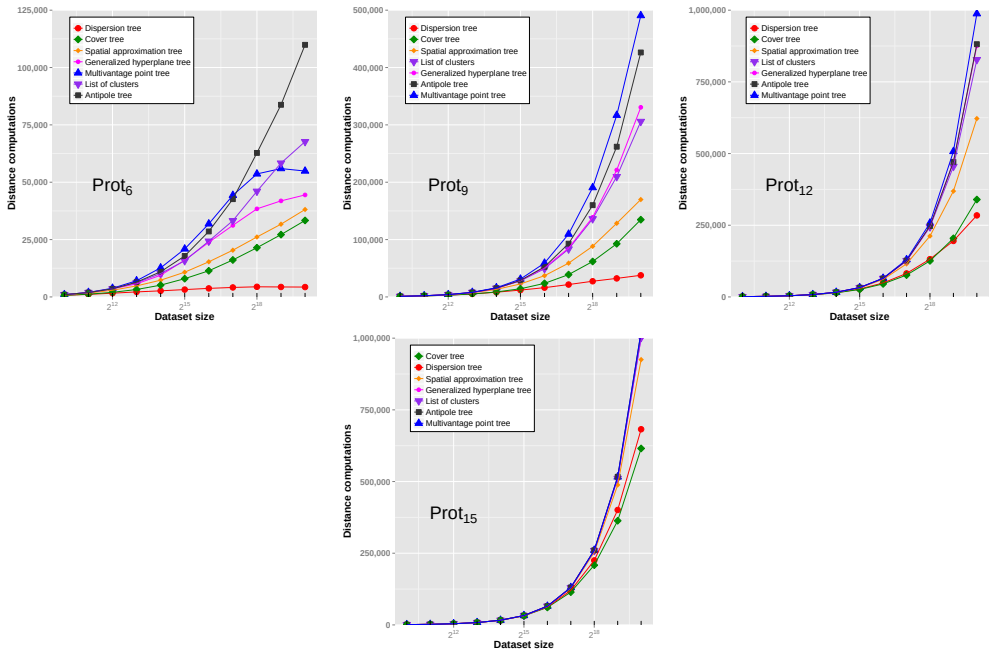


FIGURE 5. *The effects of dataset size on* 1*-NN search times in a structured metric space.* Mean 1-NN query times on the `Prot` dataset are shown as a function of the dataset size (x-axis) and kmer length (from left to right, varying from 6 to 15).

.

0.5.7 Construction

We evaluated the construction time of the metric indexes used herein by recording the number of distance computations required to construct the data structure in the $\mathbf{Eucl}_{10}$ metric space. While dispersion trees only take $O(n^{1.5} \log n)$ time to create, when compared to other metric indexes their construction times in practice are relatively slow (Figure 9). As Figure 9 is on a log-log scale, construction times that scale as n^x will appear as a straight line. Both the list of clusters and dispersion trees scale similarly, with the constants being higher for dispersion trees than the list of clusters. This is unsurprising, as the list of clusters has $\Theta(n^{1.5})$ construction time with the construction parameters used here. The other data structures exhibit construction times that appear asymptotically slower than n^x, though for some of these structures the construction time is also a function of the dimension of the space (e.g., cover trees), and the parameters to their construction (e.g., list of clusters). Also note that the construction of the index needs to only happen once, and the investment in a larger construction time can become economical given that the number of queries is sufficiently large.

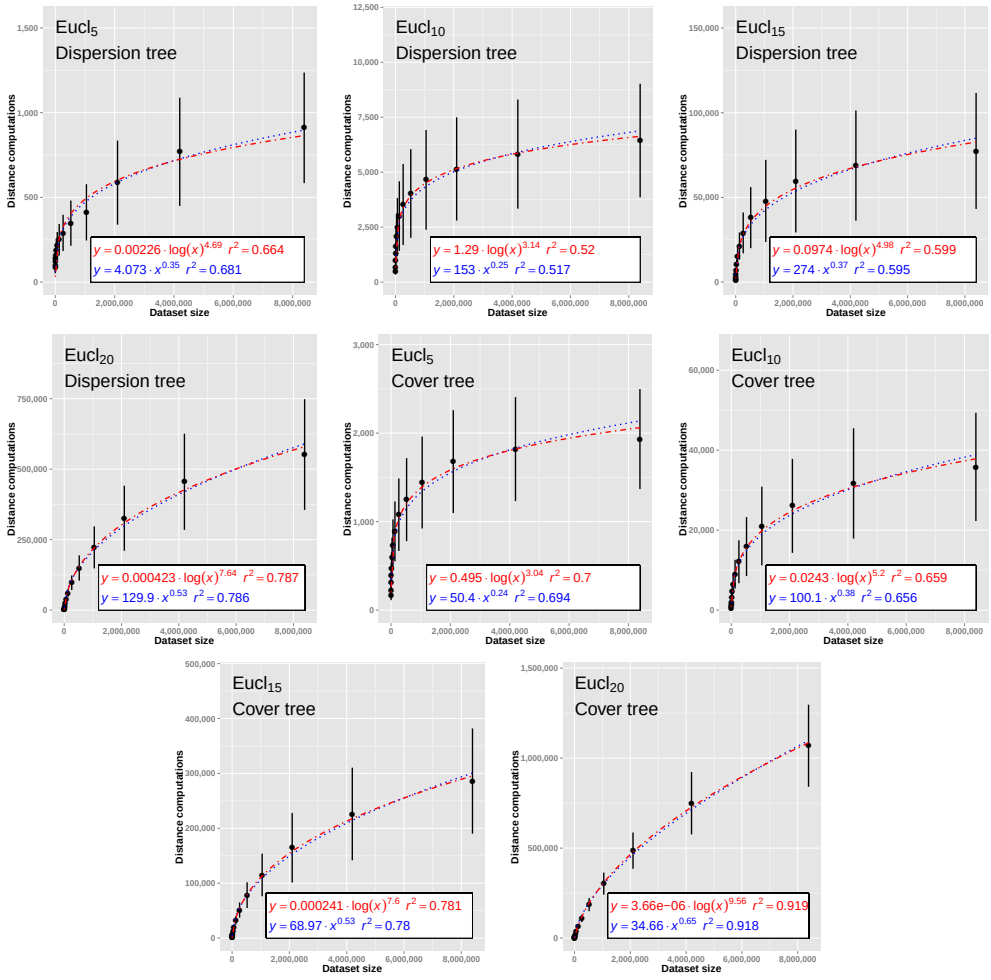


FIGURE 6. *Empirical estimates of the effects of dataset size on* 1*-NN search times in an unstructured metric space.* Mean 1-NN query times on the `Eucl` dataset are shown as a function of the dataset size (x-axis) and kmer length (from left to right, varying from 5 to 20). The plots are on a linear scale, with the mean search times ± 1 standard deviation shown in black. Nonlinear regressions were fit to the data for two functional forms, shown in red and blue, along with the coefficient of determination (r^2) for each fit. The top plots are for a *dispersion tree*, while the bottom plots are for a *cover tree*

.

0.5.8 Center Qualities

Another important contribution from this body of work is our center finding heuristic. We took 10 random subsets of the **Eucl** and **Prot** datasets of sizes ranging from 2^{10} to 2^{17}, and computed our $O(n \log n)$ time heuristic for a *MinSum* center (setting

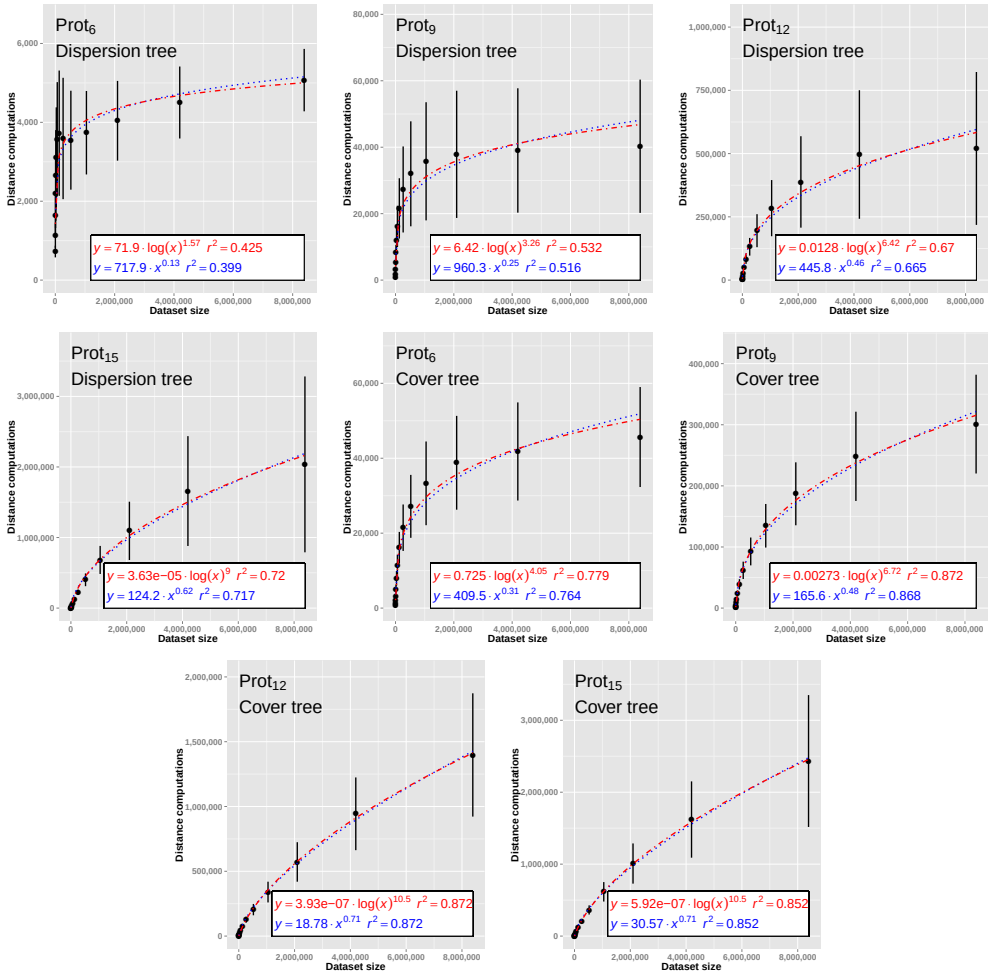


FIGURE 7. *Empirical estimates of the effects of dataset size on* 1*-NN search times in a structured metric space.* Mean 1-NN query times on the `Prot` dataset are shown as a function of the dataset size (x-axis) and kmer length (from left to right, varying from 6 to 15). The plots are on a linear scale, with the mean search times ± 1 standard deviation shown in black. Nonlinear regressions were fit to the data for two functional forms, shown in red, and blue, along with the coefficient of determination (r^2) for each fit. The top plots are for a *dispersion tree*, while the bottom plots are for a *cover tree*

.

constants r to 3 and s to 15, respectively), and we used an exhaustive search to count the number of centers that were strictly better than the center found by our heuristic. We repeated this process 10 times, and report the median and the max (worst-case) number of better centers (Figure 8). For the **DNA** dataset we used

the same experimental design, however to preserve the structure of the metric space we uses chromosomes 1 through 10 to generate our 10 replicates. Across all metric spaces, the median number of better centers remained small across both structured and unstructured metric spaces, and it appeared to be invariant to the size of the dataset as well as the kmer size / dimension (Figure 8). The worst-case behavior across the 10 replicates, while several times larger than the median, exhibits the same scaling trends as the median, suggesting that this procedure is an excellent way of finding medoids (or perhaps similar objective functions) in practice, especially as the exact solution to the `MinSum` problem is $\Theta(n^2)$.

0.6 Conclusions

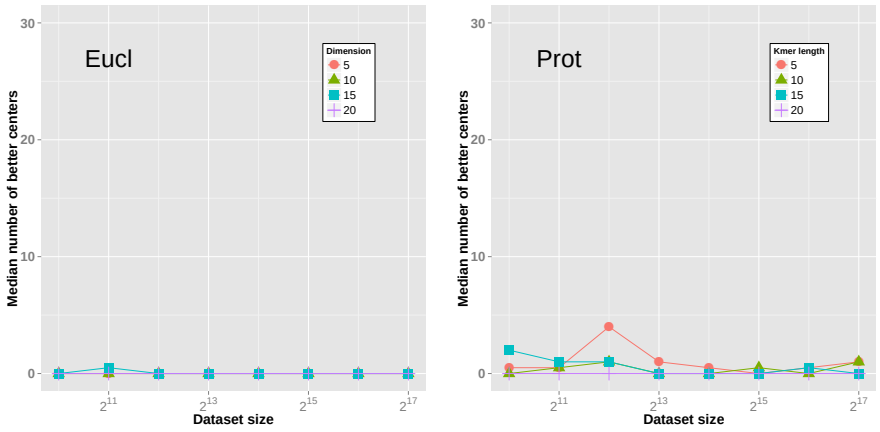
In this paper we introduce a data structure for exact k-NN search in metric spaces called a dispersion tree. Unlike all other metric space indexes, dispersion trees use a non-trivial objective function to merge subtrees in a bottom-up fashion in sub-quadratic time. This leads to an index of the metric space that is especially dense, which begets search times that are fast in medium- to high-dimensional settings and in large datasets. Our results suggest that the bottom-up construction, coupled with the appropriate objective function, leads to greatly reduced search times, and it introduces this mode of construction as a novel way to create dense indexes of high-dimensional spaces.

0.7 Further research

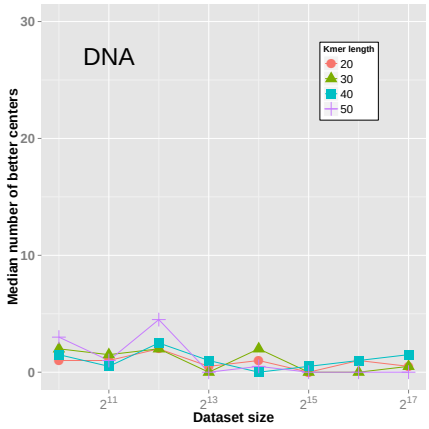
While dispersion trees have fast search times, their construction times are relatively slower than other metic indexes. Further, dispersion trees are only bottom-up with respect to our single-pass partitioning scheme. This leads us to posit that true-bottom up procedures that evaluate all points simultaneously may be a promising avenue for future inquiry. Similarly, metric index construction that operates in $O(n \log^x n)$ time for a dataset of size n may allow bottom-up construction procedures such as these to scale to much larger datasets.

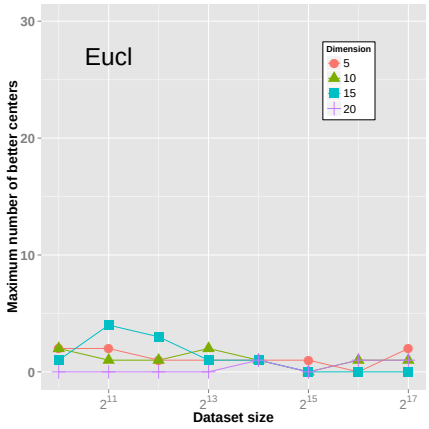
0.8 Acknowledgments

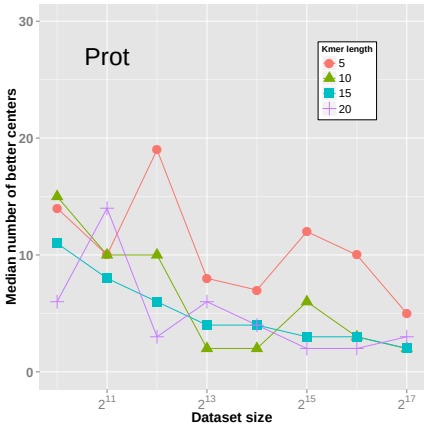
This work was supported by the National Science Foundation Graduate Research Fellowship Grant DGE-1143953 for AEW.

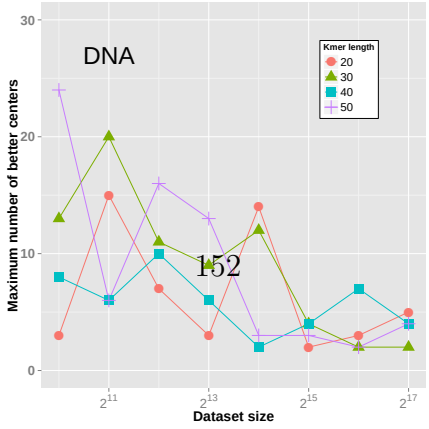

Eucl
Dimension
5
10
15
20
Median number of better centers
Dataset size

Prot
Kmer length
5
10
15
20
Median number of better centers
Dataset size


DNA
Kmer length
20
30
40
50
Median number of better centers
Dataset size


Eucl
Dimension
5
10
15
20
Maximum number of better centers
Dataset size


Prot
Kmer length
5
10
15
20
Median number of better centers
Dataset size


DNA
Kmer length
20
30
40
50
Maximum number of better centers
Dataset size

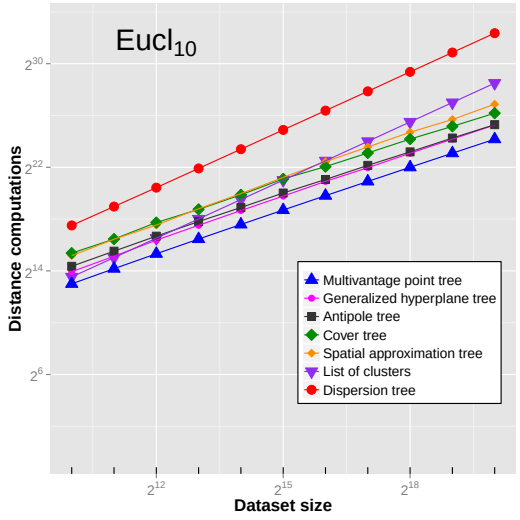


FIGURE 9. *The effects of dataset size on construction times.* The plot shows the construction time as a function of the dataset size for $\texttt{Eucl}_{10}$. This plot is on a log-log scale, which means that construction times that scale as n^x will appear as a straight line with slope x.

.

References

[1] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998.

[2] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd international conference on Machine learning*, pages 97–104. ACM, 2006.

[3] T. Bozkaya and M. Ozsoyoglu. Indexing large metric spaces for similarity search queries. *ACM Transactions on Database Systems (TODS)*, 24(3):361–404, 1999.

[4] S. Brin. Near neighbor search in large metric spaces. *Proceedings of the 21th International Conference on Very Large Data Bases (VLDB'95)*, pages 574–584, 1995.

[5] W. A. Burkhard and R. M. Keller. Some approaches to best-match file searching. *Communications of the ACM*, 16(4):230–236, 1973.

[6] D. Cantone, A. Ferro, A. Pulvirenti, D. R. Recupero, and D. Shasha. Antipole tree indexing to support range search and k-nearest neighbor search in metric spaces. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):535–550, 2005.

[7] E. Chávez and G. Navarro. An effective clustering algorithm to index high dimensional metric spaces. In *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on*, pages 75–86. IEEE, 2000.

[8] P. Ciaccia and M. Patella. Pac nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces. In *Data Engineering, 2000. Proceedings. 16th International Conference on*, pages 244–255. IEEE, 2000.

[9] K. L. Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-neighbor methods for learning and vision: theory and practice*, pages 15–59, 2006.

[10] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM, 2004.

[11] K. Figueroa, G. Navarro, and E. Chávez. Metric spaces library, 2007. Available at `http://www.sisap.org/Metric_Space_Library.html`.

[12] J. Gan, J. Feng, Q. Fang, and W. Ng. Locality-sensitive hashing scheme based on dynamic collision counting. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 541–552. ACM, 2012.

[13] G. R. Hjaltason and H. Samet. Incremental similarity search in multimedia databases. Technical report, 2000.

[14] D. Knuth. *The art of computer programming*. Addison-Wesley, 1973.

[15] Y. Liu, J. Cui, Z. Huang, H. Li, and H. T. Shen. Sk-lsh: An efficient index structure for approximate nearest neighbor search. *Proceedings of the VLDB Endowment*, 7(9), 2014.

[16] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li. Multi-probe lsh: efficient indexing for high-dimensional similarity search. In *Proceedings of the 33rd international conference on Very large data bases*, pages 950–961. VLDB Endowment, 2007.

[17] M. L. Micó, J. Oncina, and E. Vidal. A new version of the nearest-neighbour approximating and eliminating search algorithm (aesa) with linear preprocessing time and memory requirements. *Pattern Recognition Letters*, 15(1):9–17, 1994.

[18] G. Navarro. Searching in metric spaces by spatial approximation. *The VLDB Journal*, 11(1):28–46, 2002.

[19] R. Panigrahy. Entropy based nearest neighbor search in high dimensions. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1186–1195. ACM, 2006.

[20] H. Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[21] H. T. Shen, X. Zhou, and A. Zhou. An adaptive and dynamic dimensionality reduction method for high-dimensional indexing. *The VLDB Journal–The International Journal on Very Large Data Bases*, 16(2):219–234, 2007.

[22] Y. Tao, K. Yi, C. Sheng, and P. Kalnis. Quality and efficiency in high dimensional nearest neighbor search. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 563–576. ACM, 2009.

[23] J. K. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information processing letters*, 40(4):175–179, 1991.

[24] E. Vidal Ruiz. An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters*, 4(3):145–157, 1986.

[25] P. N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*, pages 311–321. Society for Industrial and Applied Mathematics, 1993.