Algorithms for Genomics and Genetics: Compression-Accelerated Search and Admixture Analysis

by

Po-Ru Loh

B.S., California Institute of Technology (2007)

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

Author . . . .
Department of Mathematics
April 23, 2013

Certified by . . . .
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by . . . .
Michel X. Goemans
Chairman, Applied Mathematics Committee

Algorithms for Genomics and Genetics: Compression-Accelerated Search and Admixture Analysis

by

Po-Ru Loh

Submitted to the Department of Mathematics
on April 23, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Rapid advances in next-generation sequencing technologies are revolutionizing genomics, with data sets at the scale of thousands of human genomes fast becoming the norm. These technological leaps promise to enable corresponding advances in biology and medicine, but the deluge of raw data poses substantial mathematical, computational and statistical challenges that must first be overcome. This thesis consists of two research thrusts along these lines. First, we propose an algorithmic framework, "compressive genomics," that accelerates bioinformatic computations through analysis-aware compression. We demonstrate this methodology with proof-of-concept implementations of compression-accelerated search (CaBLAST and CaBLAT). Second, we develop new computational tools for investigating population admixture, a phenomenon of importance in understanding demographic histories of human populations and facilitating association mapping of disease genes. Our recently released *ALDER* and *MixMapper* software packages provide fast, sensitive, and robust methods for detecting and analyzing signatures of admixture created by genetic drift and recombination on genome-wide, large-sample scales.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Acknowledgments

First, many thanks are due to my thesis adviser, Bonnie Berger, who introduced me to computational biology, guided me throughout my research, and always believed in me and my work. I also wish to especially thank David Reich and Nick Patterson, who welcomed me into the field of population genetics, and whose guidance has been essential to my work in that area. Outside of computational biology, I am grateful as well to Alan Edelman for his mentorship and many fun discussions about parallel computing and random matrices.

My graduate studies were generously funded by NDSEG and NSF graduate fellowships, NIH training grant 5T32HG004947-04, and the Simons Foundation.

I am thankful to my main collaborators on the work presented in this thesis, Mark Lipson and Michael Baym; to fellow members of the Berger group, especially George Tucker, Luke Hutchison, Michael Schnall-Levin, and Leonid Chindelevitch, from whom I have learned much; and to Patrice Macaluso, for her administrative support and kindness.

Finally, I wish to thank the people who have contributed immeasurably to my graduate experience: the Sidney-Pacific Graduate Community, especially housemasters Roger and Dottie Mark, Roland Tang and Annette Kim, for their wisdom and constant encouragement; many close friends in the MIT Graduate Christian Fellowship, whose support I deeply appreciate; and my family, whose guidance and care from the beginning til now made all this possible.

Contents

Introduction

Rapid advances in next-generation sequencing technologies are revolutionizing genomics. In the decade since the publication of the first draft of the human genome (Lander et al., 2001; Venter et al., 2001)—a 10-year, $400-million effort—new methods of image analysis, chemistry and engineering (Kircher and Kelso, 2010) have been developed that sequence a human genome in one week for less than $10,000. Data sets at the scale of thousands of human genomes are fast becoming the norm (The 1000 Genomes Project Consortium, 2012).

These technological leaps hold the promise of enabling corresponding advances in biology and medicine, but the deluge of raw data poses substantial mathematical, computational and statistical challenges that must first be overcome (Kahn, 2011; Gross, 2011; Huttenhower and Hofmann, 2010; Schatz et al., 2010). This thesis addresses such problems on two levels, (i) developing a general algorithmic framework to mitigate increasingly demanding computational requirements for genomic data analysis (Loh et al., 2012); and (ii) developing robust, efficient statistical methods that interpret genetic data to reveal insights about admixed human populations (Lipson et al., 2012; Loh et al., 2013).

Compressive genomics

Despite the rapid growth of genomic data sets—far outpacing increases in available computing power—most data now being collected is highly redundant. For example, human genomes differ on average by only 0.1% (Venter et al., 2001), so 1,000 human genomes only contain roughly twice the unique information of one genome. Many compression algorithms have been developed to increase storage and transmission efficiency (Grumbach and Tahi, 1994; Chen et al., 2002; Christley et al., 2009; Brandon et al., 2009; Mäkinen et al., 2010; Kozanitis et al., 2011; Hsi-Yang Fritz et al., 2011), but these techniques require decompression before computational analysis and thus do not mitigate the computational bottleneck.

In contrast, the "compressive genomics" framework that we have recently proposed (Loh et al., 2012) harnesses redundancy among genomes to achieve computational acceleration (in addition to data compression) by storing genomes in a compressed format that respects the structure of similarities and differences important for analysis. Once such a compressed library has been created, it can be analyzed in an amount of time proportional to its compressed size, rather than having to reconstruct the full data set every time one wishes to query it.

This approach fundamentally changes the scaling behavior of algorithmic analyses and will be essential in addressing the computational challenges now arising in the age of "big data" biology. The prototype compressive search algorithm we present in Chapter 1 represents a first step in the direction of compressive genomics, opening many avenues for further study.

Admixture inference

The explosion in genetic data has enabled significant advances in understanding the demographic histories of human populations. In particular, admixture between populations has left genetic traces in present-day populations (Wall et al., 2009; Reich et al., 2009; Green et al., 2010; Gravel et al., 2011; Pugach et al., 2011; Patterson et al., 2012) that evince past migrations and facilitate association mapping of disease genes (Patterson et al., 2004; Pasaniuc et al., 2011). Admixture must also be taken into account when correcting for population stratification among cases and controls in genome-wide association studies (Price et al., 2006; Tian et al., 2008).

Many methods have been developed to investigate admixture events using large-scale SNP data sets; some of the most popular, such as STRUCTURE (Pritchard et al., 2000) and principal component analysis (PCA) (Patterson et al., 2006), use clustering algorithms to visually identify admixed populations. Gene flow parameters can also be studied quantitatively by measuring signals of admixture from two distinct genetic processes: drift and recombination. Drift-based approaches model allele frequency divergences among populations caused by random sampling, allowing insight into admixture sources and proportions using likelihood-based models (Chikhi et al., 2001; Wang, 2003; Sousa et al., 2009; Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011) or moment statistics (Reich et al., 2009; Green et al., 2010; Patterson et al., 2012; Pickrell and Pritchard, 2012).

In contrast, recombination-based approaches harness the fact that chromosomes of admixed individuals contain continuous blocks inherited from each ancestral population that break down through successive generations. Local ancestry methods (Tang et al., 2006; Sankararaman et al., 2008; Price et al., 2009; Lawson et al., 2012) explicitly infer these blocks, enabling estimation of the age of admixture from their length distribution (Pool and Nielsen, 2009; Pugach et al., 2011; Gravel, 2012). Recent work has also shown the utility of LD statistics for date inference (Moorjani et al., 2011; Patterson et al., 2012).

The drift and recombination signals are complementary; here we extend inference methods that analyze each.

Admixture inference using linkage disequilibrium

In Chapter 2 we comprehensively develop the technique of using linkage disequilibrium (LD) in admixed populations as a source of information about population history. Linkage disequilbrium has recently emerged as a powerful signal for studying history—especially for studying migrations and inter-population gene flow—but previous methods have been somewhat rudimentary. Our work makes major new advances in both the theory and application of LD-based admixture analysis:

1. **Careful mathematical development of a novel weighted LD statistic.** The manuscript explores—in much greater depth than previous work—the exponential decay of admixture-induced LD as a function of genetic distance, applying techniques from probability theory and mathematical biology.

2. **Dating admixture 1000 times faster than previous methods.** The new method speeds up computation by a thousand-fold by applying an algebraic manipulation to recast the problem in a form that can be computed using a fast Fourier transform (FFT).

3. **A novel statistical test for admixture based on LD.** The paper shows how LD can be used as the basis of a novel three-population test for admixture and applies this test to detect admixture signals that other methods cannot, for example in Sardinians, Mbuti Pygmies, and Japanese.

4. **Novel methods for estimating other admixture parameters.** The manuscript describes how weighted LD can be used not only to estimate dates of admixture but also to infer mixture proportions and phylogenetic relationships.

We have implemented this methodology in a versatile software package, *ALDER* (Loh et al., 2013), that we anticipate will be a useful tool for a wide range of geneticists working on human history as well as the genetics of other species.

Admixture inference using moment statistics

In Chapter 3 we present *MixMapper*, a new method for investigating population history that identifies and models admixed populations using allele frequency moment statistics (Lipson et al., 2012). We demonstrate the power of *MixMapper* by applying it to worldwide samples from the Human Genome Diversity Project (HGDP). Surprisingly, *MixMapper* is able to detect for the first time a signal of ancient admixture in all European HGDP populations, altering the current understanding of the history of admixture in Europe. In particular, Sardinians and Basques have previously been widely viewed as unadmixed descendants of Neolithic farmers, but *MixMapper* identifies evidence of admixture in these populations, which we convincingly validate using multiple lines of reasoning.

Methodologically, *MixMapper* has two key features that distinguish it from existing approaches:

1. **Automatic identification of admixed populations and sources of gene flow.** Nearly all previous model-based admixture inference methods require the user to specify a pre-determined phylogeny, which can sometimes be incorrect. In contrast, *MixMapper* computationally explores all possible phylogenies and determines the topological relationships that best explain the data. This ability makes it particularly effective for understanding relationships involving unknown ancestral mixing populations (e.g., ancient European populations).

2. **Fast, interactive, and scalable computation.** Following modest setup time of roughly an hour for a given data set, each admixture fit in *MixMapper* takes only seconds. This time scale is fast enough that the investigator can interactively explore the robustness of inferences to perturbations of the model assumptions, promoting greater understanding of the statistical significance of reported results. Also, as opposed to likelihood-based methods, which can typically only handle a few populations,

the moment-based approach *MixMapper* employs can efficiently accommodate dozens of populations.

We anticipate that *MixMapper* will be an important addition to the toolkit presently available for studying the evolutionary genetics of humans and other species.

Chapter 1

Compression-Accelerated Search

In the past two decades, genomic sequencing capabilities have increased exponentially (Lander et al., 2001; Venter et al., 2001; Kircher and Kelso, 2010), outstripping advances in computing power (Kahn, 2011; Gross, 2011; Huttenhower and Hofmann, 2010; Schatz et al., 2010; The 1000 Genomes Project Consortium, 2012). Extracting new insights from the data sets currently being generated will require not only faster computers, but also smarter algorithms. However, most genomes currently sequenced are highly similar to ones already collected (Stratton, 2008); thus, the amount of new sequence information is growing much more slowly.

Here we show that this redundancy can be exploited by compressing sequence data in such a way as to allow direct computation on the compressed data using methods we term 'compressive' algorithms. This approach reduces the task of computing on many similar genomes to only slightly more than that of operating on just one. Moreover, its relative advantage over existing algorithms will grow with the accumulation of genomic data. We demonstrate this approach by implementing compressive versions of both the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) and the BLAST-Like Alignment Tool (BLAT) (Kent, 2002), and we emphasize how compressive genomics will enable biologists to keep pace with current data.[1]

1.1 A changing environment

Successive generations of sequencing technologies have increased the availability of genomic data exponentially. In the decade since the publication of the first draft of the human genome (a 10-year, $400-million effort (Lander et al., 2001; Venter et al., 2001)), technologies (Kircher and Kelso, 2010) have been developed that can be used to sequence a human genome in 1 week for less than $10,000, and the 1000 Genomes Project is well on its way to building a library of over 2,500 human genomes (The 1000 Genomes Project Consortium, 2012).

These leaps in sequencing technology promise to enable corresponding advances in biology and medicine, but this will require more efficient ways to store, access and analyze large genomic data sets. Indeed, the scientific community is becoming aware of the fundamental

[1] The material in this chapter previously appeared in the July 2012 issue of *Nature Biotechnology* as "Compressive Genomics" by Po-Ru Loh, Michael Baym, and Bonnie Berger (Loh et al., 2012).

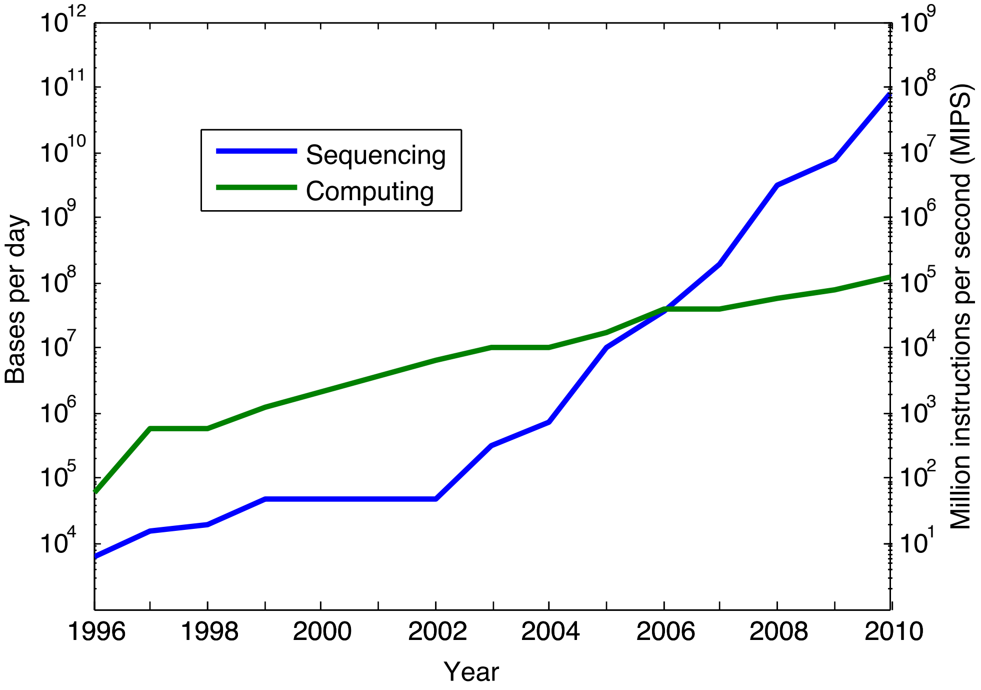


Figure 1.1. Sequencing capabilities vs. computational power from 1996-2010. Sequencing capabilities are doubling approximately every four months, whereas processing capabilities are doubling approximately every eighteen. (Data adapted with permission from Kahn (2011).)

challenges in analyzing such data (Kahn, 2011; Gross, 2011; Huttenhower and Hofmann, 2010; Schatz et al., 2010). Difficulties with large data sets arise in settings in which one analyzes genomic sequence libraries, including finding sequences similar to a given query (e.g., from environmental or medical samples) or finding signatures of selection in large sets of closely related genomes.

Currently, the total amount of available genomic data is increasing approximately tenfold every year, a rate much faster than Moore's Law for computational processing power (Fig. 1.1). Any computational analysis, such as sequence search, that runs on the full genomic library—or even a constant fraction thereof—scales at least linearly in time with respect to the size of the library and therefore effectively grows exponentially slower every year. If we wish to use the full power of these large genomic data sets, then we must develop new algorithms that scale sublinearly with data size (that is, those that reduce the effective size of the data set or do not operate on redundant data).

1.2 Sublinear analysis and compressed data

To achieve sublinear analysis, we must take advantage of redundancy inherent in the data. Intuitively, given two highly similar genomes, any analysis based on sequence similarity that is performed on one should have already done much of the work toward the same analysis on the other. We note that although efficient algorithms, such as BLAST (Altschul et al., 1990), have been developed for individual genomes, large genomic libraries have additional structure: they are highly redundant. For example, as human genomes differ on average by only 0.1% (Venter et al., 2001), 1,000 human genomes contain less than twice the unique information of one genome. Thus, although individual genomes are not very compressible (Grumbach and Tahi, 1994; Chen et al., 2002), collections of related genomes are extremely compressible (Christley et al., 2009; Brandon et al., 2009; Mäkinen et al., 2010; Kozanitis et al., 2011).

This redundancy among genomes can be translated into computational acceleration by storing genomes in a compressed format that respects the structure of similarities and differences important for analysis. Specifically, these differences are the nucleotide substitutions, insertions, deletions and rearrangements introduced by evolution. Once such a compressed library has been created, it can be analyzed in an amount of time proportional to its compressed size, rather than having to reconstruct the full data set every time one wishes to query it.

Many algorithms exist for the compression of genomic data sets purely to reduce the space required for storage and transmission (Grumbach and Tahi, 1994; Chen et al., 2002; Christley et al., 2009; Brandon et al., 2009; Kozanitis et al., 2011; Hsi-Yang Fritz et al., 2011). Hsi-Yang Fritz et al. (2011) provide a particularly instructive discussion of the concerns involved. However, existing techniques require decompression before computational analysis. Thus, although these algorithms enable efficient data storage, they do not mitigate the computational bottleneck: the original uncompressed data set must be reconstructed before it can be analyzed.

There have been efforts to accelerate exact search through indexing techniques (Mäkinen et al., 2010; Deorowicz and Grabowski, 2011). Although algorithms—such as Maq (Li et al., 2008a), Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) and Bowtie (Langmead et al., 2009)—already can map short resequencing reads to a few genomes quite well, compressive techniques will be extremely useful in the case of matching reads of unknown origin to a large database (say, in a medical or forensic context). Search acceleration becomes harder when one wishes to perform an inexact search (e.g., BLAST and BLAT) because compression schemes in general do not allow efficient recovery of the similarity structure of the data set.

As proof of principle for the underlying idea of compressive genomics, we present model compressive algorithms that run BLAST and BLAT in time proportional to the size of the nonredundant data in a genomic library. We chose BLAST for a primary demonstration because it is widely used and also the principal means by which many other algorithms query large genomic data sets; thus any improvement to BLAST will immediately improve various analyses on large genomic data sets. Furthermore, the compressive architecture for sequence search we introduce here is tied not only to BLAST but also to many algorithms (particularly those based on sequence similarity).

1.3 Compressive genomics using BLAST and BLAT

We describe versions of the widely used BLAST and BLAT algorithms that illustrate the compressive genomics paradigm. BLAST and BLAT search a genomic database to identify sequences that are similar to a given sequence. Our compressive algorithms have two phases: (i) compressing the database and (ii) searching the compressed data (Fig. 1.2). The compression phase can be realized by various schemes. We used an approach based on edit script compression. The search phase can be implemented using nearly any existing search algorithm. We show the modularity of our approach by implementing compressive BLAST and BLAT search algorithms that can operate on the same compressed database.

1.3.1 Database compression

To compress data, we store only the differences between similar sequence fragments, rather than the complete, highly redundant sequences themselves. We implement this approach by scanning the nucleotide database and identifying sequence fragments sufficiently similar to previously seen fragments. Once identified, each fragment is replaced with a link to the original sequence and a compact list of differences. By default, we consider only fragments 300 base pairs or longer with at least 85% identity to previous fragments (Appendix A). The initial data-compression phase only needs to be done once, and the compressed database can be updated incrementally if new data are added. This approach substantially reduces the storage required for many genomes (Fig. 1.3a).

The exact output of compression is dependent on the order in which the uncompressed data are examined; however, changing the order in which genomes are added to the library does not substantially affect the database size, compression speed, search speed or search accuracy (data not shown). For example, using our compressive BLAST algorithm, accuracy to hits in the first genome added to the database was perfect, and the accuracy of all subsequent hits was <1% lower.

1.3.2 Compressive BLAST

For the search phase, we implemented a two-step variant of BLAST. First, the algorithm uses standard BLAST to search the unique data (that is, data not replaced by links during compression) with a more permissive hit threshold (E value). Second, the algorithm traces links to determine potential hit regions in the full database and examines these potential hits with the original, stricter threshold. The initial 'coarse' search runs on the compressed data without decompression, yielding a run time proportional to the size of the compressed database. The second 'fine' alignment is done by locally decompressing only the potential hit regions until either a hit is determined or the region can be ruled out.

As the coarse search has a relaxed threshold, searches that hit repeat regions will result in many more coarse hits and thus burden the full computation. In practice, we mitigate this issue by masking repeat regions. For the results presented here, we used a coarse E value threshold of 10^{-20}, and always set the BLAST database size parameter to the size of the uncompressed database (Appendix A).

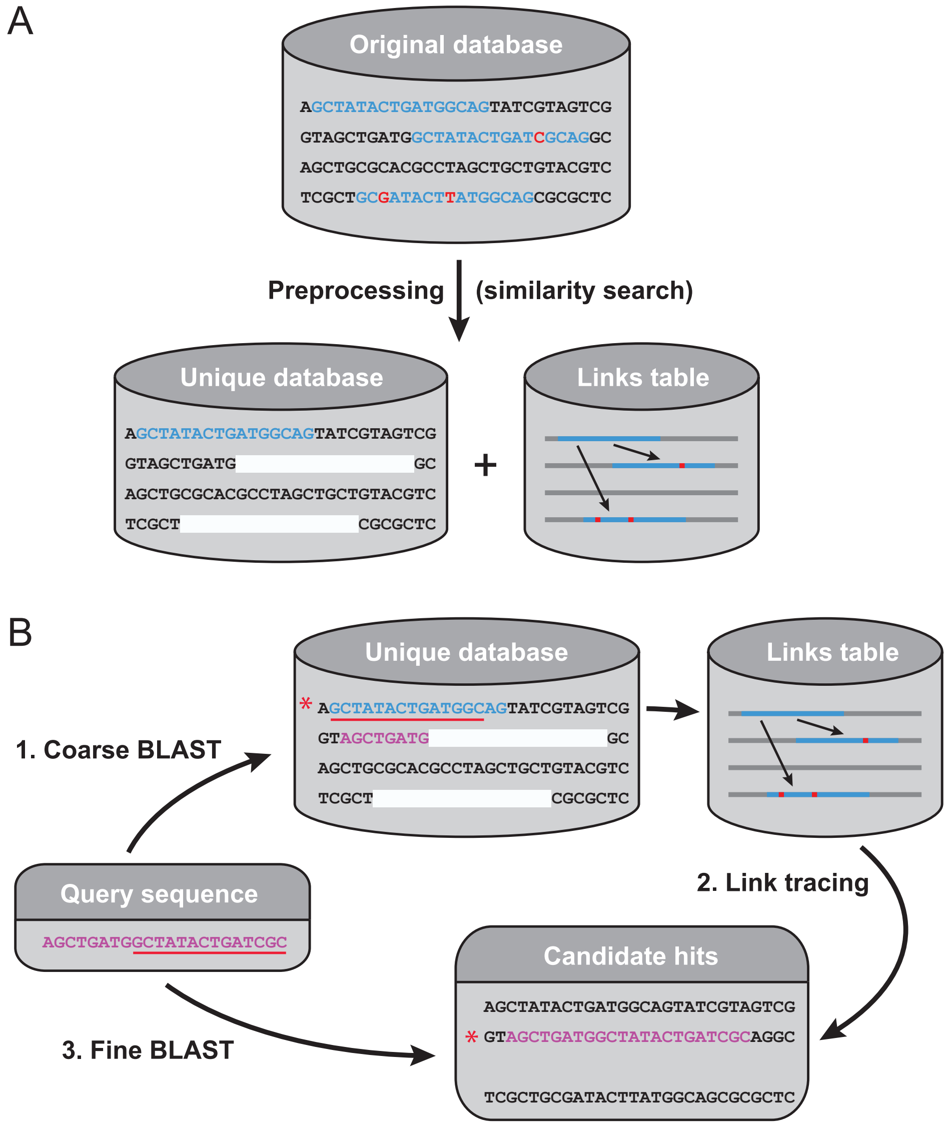

A
Original database
AGCTATACTGATGGCAGTATCGTAGTCG
GTAGCTGATGGCTATACTGATCGCAGGC
AGCTGCGCACGCCTAGCTGCTGTACGTC
TCGCTGCGATACTTATGGCAGCGCGCTC
Preprocessing
(similarity search)
Unique database
AGCTATACTGATGGCAGTATCGTAGTCG
GTAGCTGATG
GC
AGCTGCGCACGCCTAGCTGCTGTACGTC
TCGCT
CGCGCTC
+
Links table
B
Unique database
*
AGCTATACTGATGGCAGTATCGTAGTCG
GTAGCTGATG
GC
AGCTGCGCACGCCTAGCTGCTGTACGTC
TCGCT
CGCGCTC
Links table
1. Coarse BLAST
Query sequence
AGCTGATGGCTATACTGATCGC
2. Link tracing
Candidate hits
AGCTATACTGATGGCAGTATCGTAGTCG
*
GTAGCTGATGGCTATACTGATCGCAGGC
TCGCTGCGATACTTATGGCAGCGCGCTC
3. Fine BLAST

Figure 1.2. The CaBLAST algorithm. (*A*) Preprocessing: Scan through the genomic database, eliminating redundancy by identifying regions of high similarity and encoding them as links to previously seen sequences. The blue sequence fragments in the figure match perfectly, aside from a few discrepancies (red). Thus, only the first occurrence appears in the unique database; information about the other two fragments is stored in the links table. (*B*) Search: (1) BLAST against unique database with relaxed E-value threshold; (2) Recover additional candidate hits via links table; (3) BLAST against candidate hits to select final results. Here, the query (purple) matches the second original sequence, but most of the second sequence does not appear in the unique database because of a similar region in the first sequence (blue). The coarse BLAST query thus hits only the first sequence; the second sequence is recovered as a candidate after checking the links table.

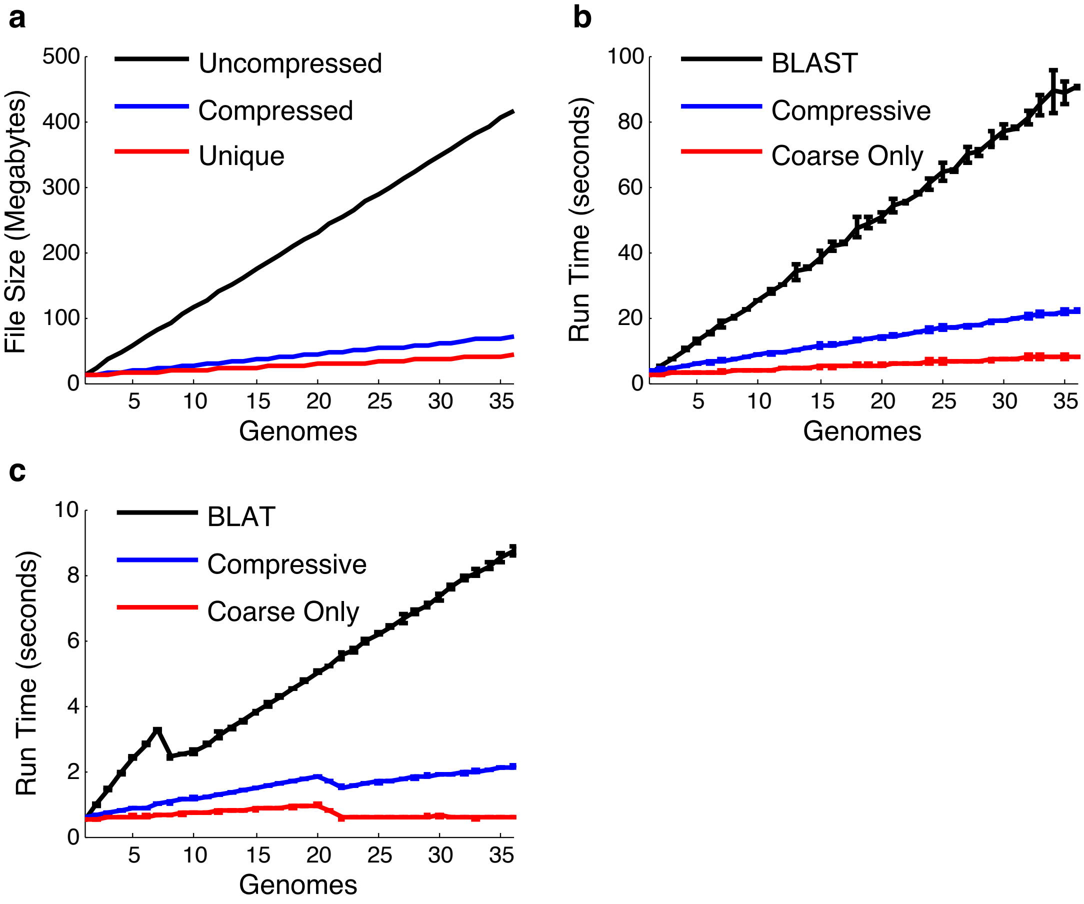


Figure 1.3. Results of compressive algorithms on up to 36 yeast genomes. (a) File sizes of the uncompressed (black), compressed with links and edits (blue), and unique sequence (red) datasets with default parameters. (b) Runtimes of BLAST (black), Compressive BLAST (blue), and the coarse search step of Compressive BLAST on the unique data (red). Error bars are the standard deviation of five runs. Runtimes are on a set of 10,000 simulated queries. , The coarse search time provides a lower bound on search time that would be achieved for query sets that generate very few hits. (c) Runtimes of BLAT (black), Compressive BLAT (blue), and the coarse search step on the unique data (red) for 10,000 queries. Error bars are the standard deviation of five runs. BLAST and BLAT are both run with default parameters. The data displayed represent differences between searches with 10,000 and 20,000 queries so as to remove the bias introduced by database construction time in BLAT. The anomalous decrease in runtime with more data at 8 uncompressed genomes or 21 compressed genomes is a repeatable behavior of BLAT with default parameters.

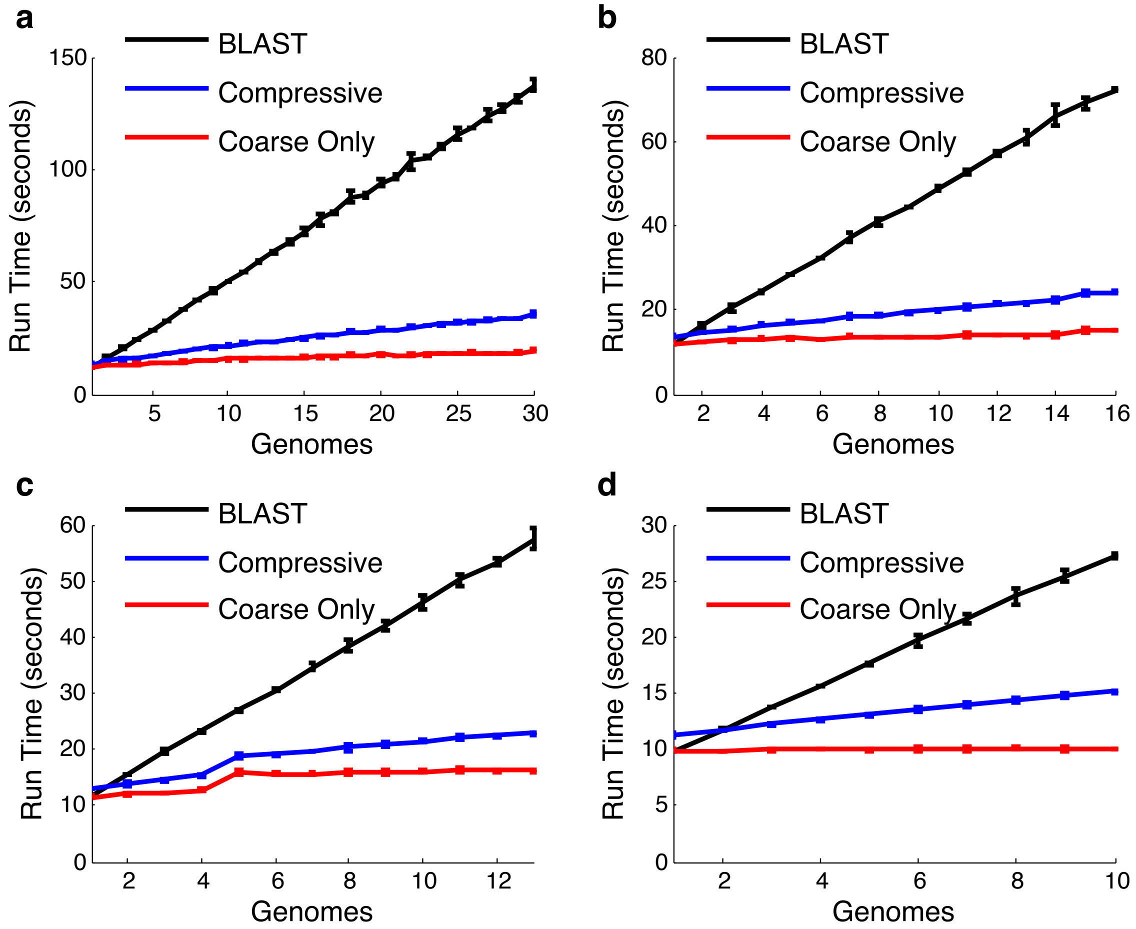


Figure 1.4. Performance of Compressive BLAST on databases of four bacterial genera using a single search set derived from the combined library of bacterial and yeast sequences. (a) Escherichia; (b) Salmonella; (c) Yersinia; (d) Brucella.

To determine whether compression yields acceleration, we compressed 36 *Saccharomyces* sp. genomes (Liti et al., 2009) (Fig. 1.3a), four sets of bacterial genera and twelve Drosophila sp. fly genomes (Tweedie et al., 2009). We simulated queries by sampling from the data set and adding mutations, producing a set of queries with many expected hits.

Compressive BLAST analysis of the yeast data set achieved a more than fourfold increase in speed with respect to a BLAST analysis. As expected, the advantage increased substantially with the number of genomes (Fig. 1.3b). We found a similar increase in speed for the microbial data sets (Figs. 1.4, 1.5 and 1.6). As our queries had many hits, the majority of the computation time (≈73% for yeast) was spent on the fine search step, whereas for queries with few hits, the coarse step alone would be a more accurate predictor of run time. We expect that much faster fine search times can be achieved with an optimized fine search algorithm; our implementation simply runs BLAST a second time on potential hit regions.

For the fly species, although we achieved a large increase in search speed for the closely related *D. melanogaster*, *D. simulans* and *D. sechellia* genomes, the gains diminished as we included more distant cousins (Table 1.1). In general, the run time of our compressive

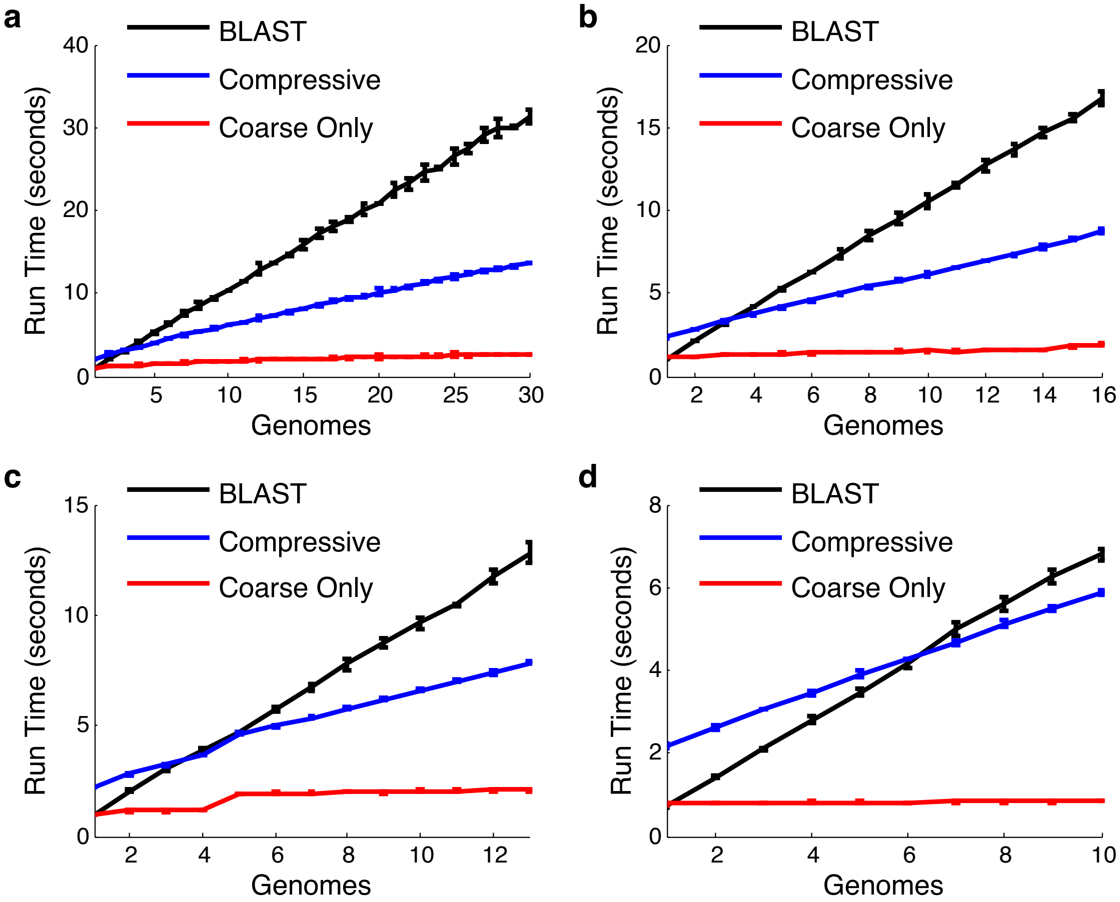


Figure 1.5. Performance of Compressive BLAST on databases of four bacterial genera using distinct search sets derived from each bacterial genus individually. Results, while still showing improvement over BLAST, are slower than the test set of Fig. 1.4, representing the increased time spent in fine search owing to increased hit rates. (a) Escherichia; (b) Salmonella; (c) Yersinia; (d) Brucella.

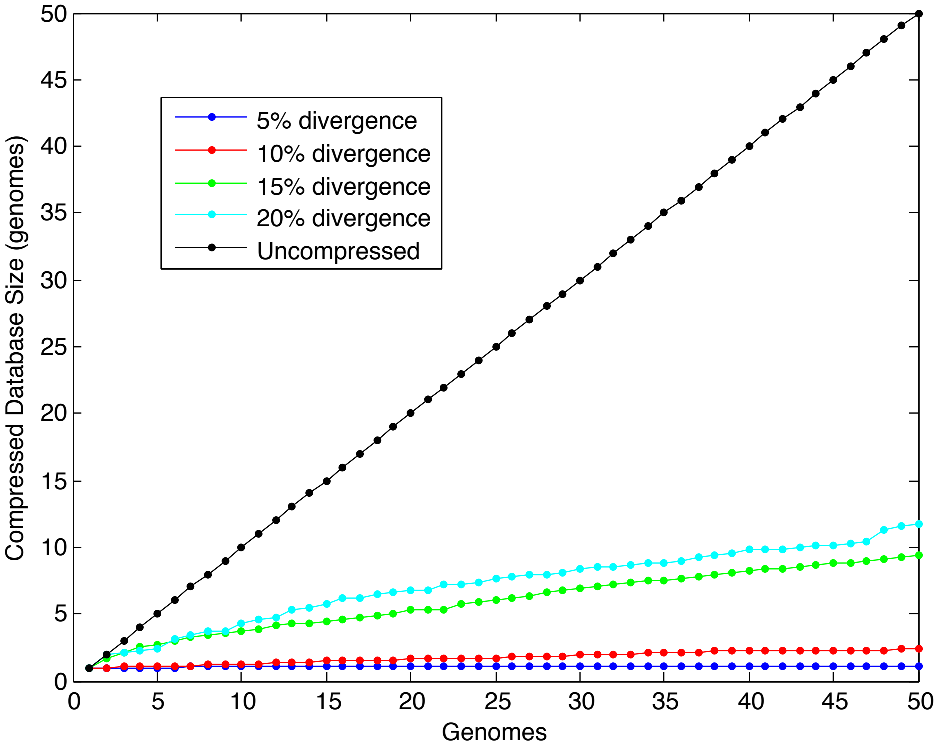


Figure 1.6. Performance of the Compressive BLAST preprocessing phase on simulated genera. Databases consist of sets of 50 simulated genomes (at 5%, 10%, 15%, and 20% divergence) generated with INDELible v1.03 (Fletcher and Yang, 2009).

Species set	BLAST runtime (sec)	Coarse E-value	CaBLAST runtime (sec)		Total % of BLAST runtime	Accuracy (%)
			Coarse BLAST	Fine BLAST		
{sim,sec}	97	1e-20	54	3.1	59	99.6
Original size: 304 Mb		1e-15	53	3.6	58	100.0
Unique size: 174 Mb (57.1%)		1e-10	55	8.0	65	100.0
{sim,sec,mel}	147	1e-20	72	4.8	53	98.9
Original size: 473 Mb		1e-15	73	6.4	54	99.7
Unique size: 235 Mb (49.7%)		1e-10	74	13.7	60	99.8
{yak,ere}	96	1e-20	67	2.5	73	99.6
Original size: 318 Mb		1e-15	67	3.2	73	99.8
Unique size: 225 Mb (70.5%)		1e-10	69	8.8	80	99.9
{sim,sec,mel,yak,ere}	243	1e-20	126	6.7	55	98.9
Original size: 792 Mb		1e-15	127	9.0	56	99.6
Unique size: 411 Mb (57.3%)		1e-10	130	20.6	62	99.8

Table 1.1. Detailed preprocessing and search results from CaBLAST runs on four sets of *Drosophila* genomes using final E-value 1e-30 and three choices of coarse E-value. CaBLAST runtime is dominated by the coarse BLAST step, achieving speedup relative to BLAST almost exactly proportional to the reduction from the original to the unique database (size ratio in parentheses). Accuracies nearly match those of BLAST; note that specificity is always 100% by virtue of the algorithm design: hits picked up by the fine search are by definition true BLAST hits.

technique scales linearly with respect to the size of the nonredundant component of the library (which we expect to be a diminishing proportion), and linearly in the number of coarse hits.

1.3.3 Compressive BLAT

To implement a compressive version of the faster BLAT algorithm, we substituted BLAT for BLAST in the coarse search step and used BLAT's local alignment algorithm for the fine search to ensure comparable results. We tested compressive BLAT on the same data as above using BLAT's `minIdentity` parameter for coarse and fine search thresholds (`minIdentity` = 80 and 90, respectively).

Our compressive approach achieved acceleration over BLAT comparable to our results from accelerating BLAST (Figs. 1.3c and 1.7). Although the coarse search step in BLAT theoretically takes a constant amount of time, in practice the running time of BLAT on a database of many genomes scales linearly with database size owing to the existence of many more 10-mer seed matches found during a search. Compression accelerates this step by allowing BLAT to rule out families of spurious hits only once. The hits produced by compressive BLAT analysis had an overall 96% accuracy and 97% specificity with respect to a BLAT analysis. The hits found by one algorithm and not the other were overwhelmingly of weak similarity. Thus, although it did not produce precisely the same hits as BLAT, compressive BLAT obtained coverage of true hits similar to the performance of BLAT.

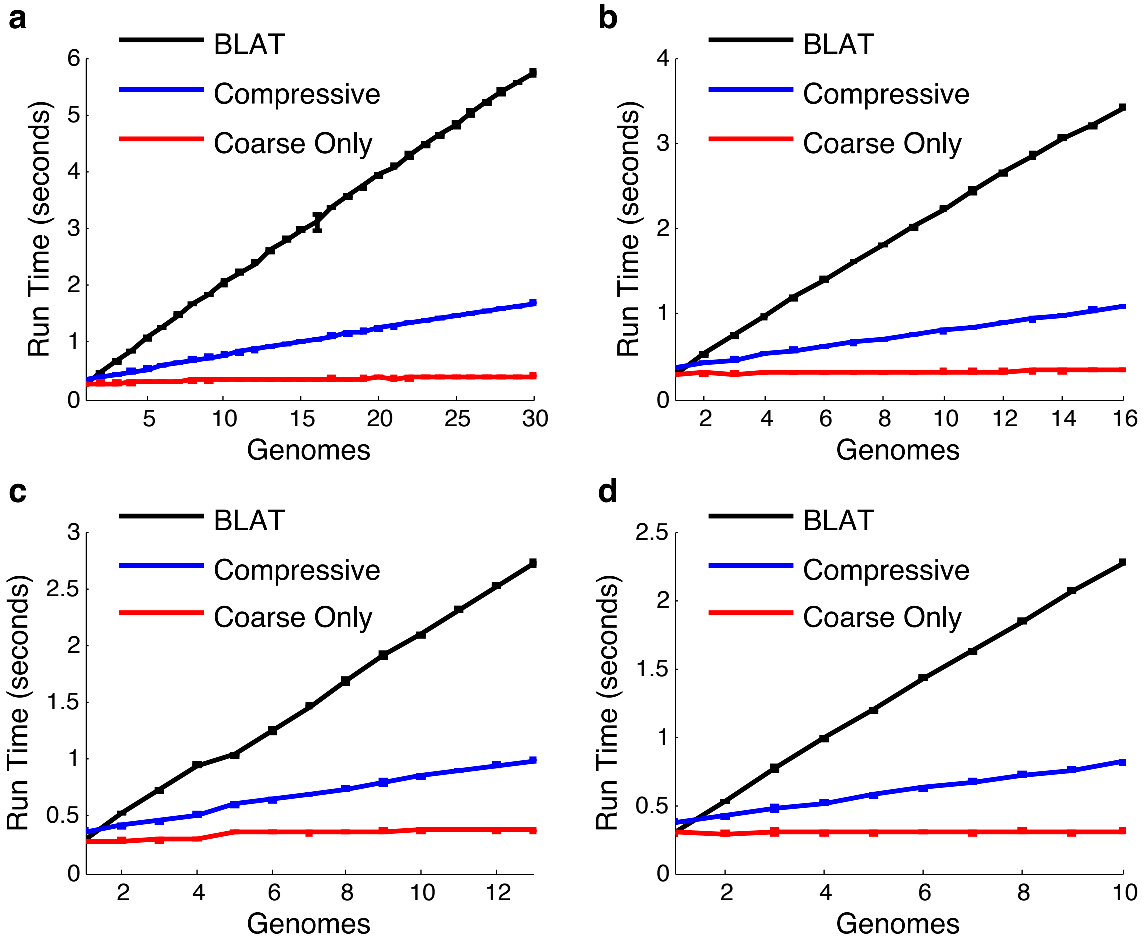


Figure 1.7. Performance of Compressive BLAT on databases of four bacterial genera using distinct search sets derived from each bacterial genus individually. Parameters are the same (default) as in the primary manuscript. (a) Escherichia; (b) Salmonella; (c) Yersinia; (d) Brucella.

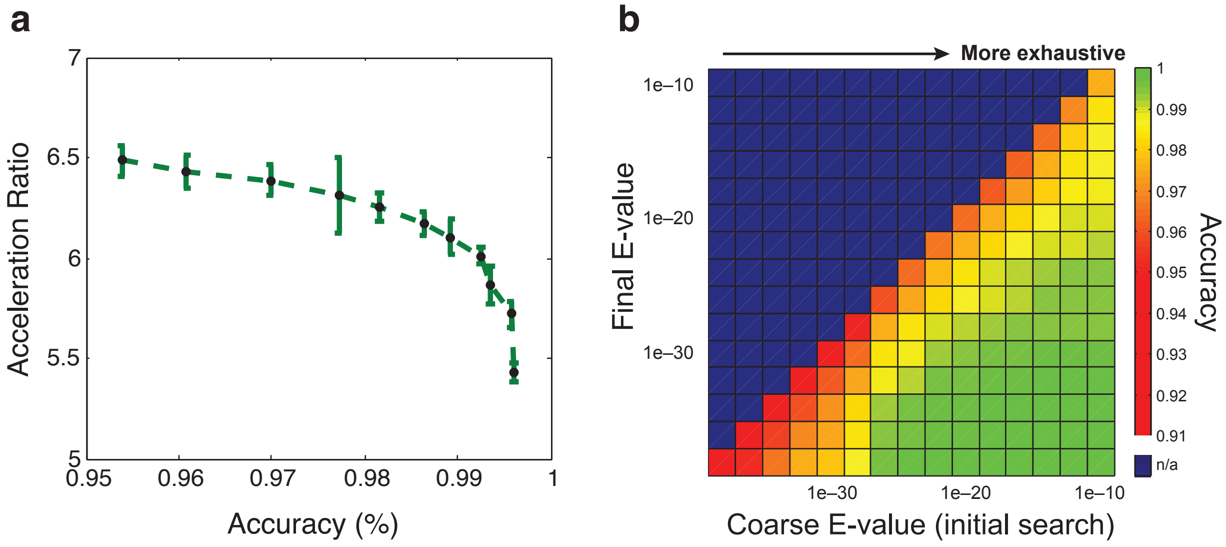


Figure 1.8. Trade-offs in Compressive BLAST. Data presented are from runs on the combined microbial data set (yeast genomes and four bacterial genera) with search queries drawn randomly from the combined library. Except where explicitly parameterized, default values for the compression threshold (85%) and coarse E-value (10^{-30}) were used. (a) Speed vs. accuracy as a function of the match identity threshold in database compression. From left to right, the points represent thresholds of 70-90%, with points every 2%. (b) Accuracy as a function of coarse and fine E-value thresholds.

1.4 Challenges of compressive algorithms

There are trade-offs to this approach. As more divergent genomes are added to a database, the computational acceleration resulting from compression decreases, although this is to be expected, as these data are less mutually informative. Although our compressive BLAST algorithm achieves over 99% sensitivity without substantial slowdown (Figs. 1.8, 1.9 and 1.10), improvements in sensitivity necessarily involve losses in speed.

There is also a trade-off between achieving optimal data compression and accuracy of analysis (Fig. 1.9a). This trade-off is fundamental to the problem of compressive algorithms for biology: in genomic analysis, one is interested in the probability of similar sequences occurring by chance rather than because of common ancestry, whereas compression ratios depend only on the absolute sequence similarity. For example, two sequences of 50% identity for over 1,000 bases are a strong BLAST hit, but admit no useful compression because the overhead would outweigh the savings. Although these two measures of sequence similarity are closely related, the difference is at the root of these trade-offs. However, sacrificing some accuracy of distant matches helps to achieve a dramatic increase in speed from compression.

As computing moves toward distributed and multiprocessor architectures, one must consider the ability of new algorithms to run in parallel. Although we expect that the primary method of parallelizing compressive genomic search algorithms will be to run queries independently, truly massive data sets will require single queries to be executed in parallel as well. In the algorithms presented here, queries can be parallelized by dividing the compressed

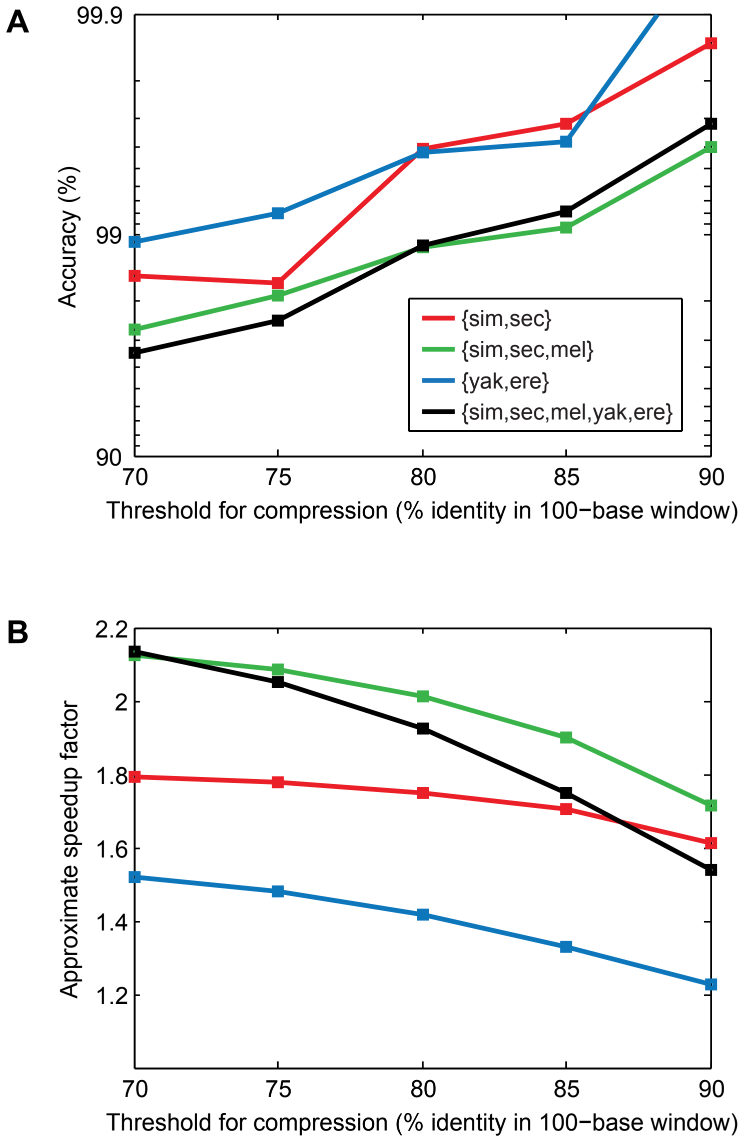


Figure 1.9. Trade-off between preprocessing compression and search accuracy for simulated queries on *Drosophila* subtrees. As the threshold required for compression is increased from 70 to 90% identity within each 100-base window, accuracy improves (*A*) while search speedup decreases (*B*).

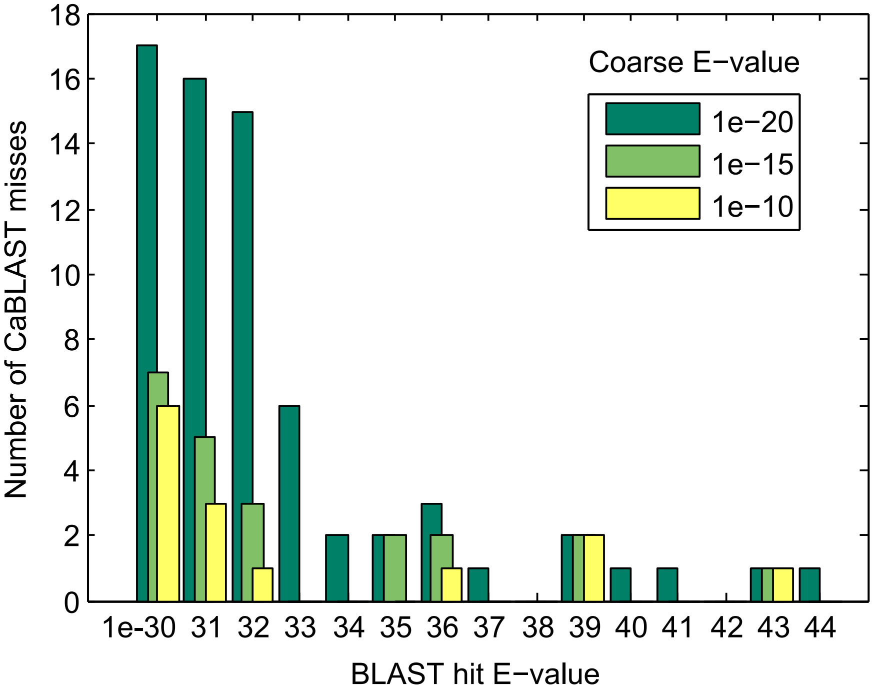


Figure 1.10. Analysis of missed BLAST hits. Ten thousand queries were run on the combined microbial dataset (yeast plus four bacterial genera) at three different coarse E-values and a fine E-value of 1e-30. The overwhelming majority of misses are at the margin of significance; in total these represent less than 0.5% of the hits.

library and link table among computer processors, although the exact gains from doing so will depend on the topology of the link graph on the uncompressed database.

To the extent that researchers restrict their analyses to small data sets (e.g., what could be generated in a single laboratory as opposed to a large sequencing center), existing noncompressive custom pipelines may be sufficiently fast in the short term. However, if one wishes to extend an analysis to a much larger corpus of sequencing data (perhaps several terabytes of raw data), noncompressive approaches quickly become computationally impractical. This is where compressive algorithms are useful for smaller research groups in addition to large centers.

1.5 Conclusions

Compressive algorithms for genomics have the great advantage of becoming proportionately faster with the size of the available data. Although the compression schemes for BLAST and BLAT that we presented yield an increase in computational speed and, more importantly, in scaling, they are only a first step. Many enhancements of our proof-of-concept implementations are possible; for example, hierarchical compression structures, which respect the phylogeny underlying a set of sequences, may yield additional long-term performance gains. Moreover, analyses of such compressive structures will lead to insights as well. As sequencing technologies continue to improve, the compressive genomic paradigm will become critical to fully realizing the potential of large-scale genomics.

Software is available at `http://cast.csail.mit.edu/`.

Chapter 2

Admixture Inference Using Linkage Disequilibrium

Long-range migrations and the resulting admixtures between populations have been important forces shaping human genetic diversity. Most existing methods for detecting and reconstructing historical admixture events are based on allele frequency divergences or patterns of ancestry segments in chromosomes of admixed individuals. An emerging new approach harnesses the exponential decay of admixture-induced linkage disequilibrium (LD) as a function of genetic distance. Here, we comprehensively develop LD-based inference into a versatile tool for investigating admixture. We present a new weighted LD statistic that can be used to infer mixture proportions as well as dates with fewer constraints on reference populations than previous methods. We define an LD-based three-population test for admixture and identify scenarios in which it can detect admixture events that previous formal tests cannot. We further show that we can uncover phylogenetic relationships among populations by comparing weighted LD curves obtained using a suite of references. Finally, we describe several improvements to the computation and fitting of weighted LD curves that greatly increase the robustness and speed of the calculations. We implement all of these advances in a software package, *ALDER*, which we validate in simulations and apply to test for admixture among all populations from the Human Genome Diversity Project (HGDP), highlighting insights into the admixture history of Central African Pygmies, Sardinians, and Japanese.[1]

2.1 Introduction

Admixture between previously diverged populations has been a common feature throughout the evolution of modern humans and has left significant genetic traces in contemporary populations (Li et al., 2008b; Wall et al., 2009; Reich et al., 2009; Green et al., 2010; Gravel et al., 2011; Pugach et al., 2011; Patterson et al., 2012). Resulting patterns of variation can provide information about migrations, demographic histories, and natural selection and can also be a valuable tool for association mapping of disease genes in admixed populations (Patterson

[1] The material in this chapter previously appeared in the April 2013 issue of *Genetics* as "Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium" by Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K. Pickrell, David Reich, and Bonnie Berger (Loh et al., 2013).

et al., 2004).

Recently, a variety of methods have been developed to harness large-scale genotype data to infer admixture events in the history of sampled populations, as well as to estimate a range of gene flow parameters, including ages, proportions, and sources. Some of the most popular approaches, such as STRUCTURE (Pritchard et al., 2000) and principal component analysis (PCA) (Patterson et al., 2006), use clustering algorithms to identify admixed populations as intermediates in relation to surrogate ancestral populations. In a somewhat similar vein, local ancestry inference methods (Tang et al., 2006; Sankararaman et al., 2008; Price et al., 2009; Lawson et al., 2012) analyze chromosomes of admixed individuals with the goal of recovering continuous blocks inherited directly from each ancestral population. Because recombination breaks down ancestry tracts through successive generations, the time of admixture can be inferred from the tract length distribution (Pool and Nielsen, 2009; Pugach et al., 2011; Gravel, 2012), with the caveat that accurate local ancestry inference becomes difficult when tracts are short or the reference populations used are highly diverged from the true mixing populations.

A third class of methods makes use of allele frequency differentiation among populations to deduce the presence of admixture and estimate parameters, either with likelihood-based models (Chikhi et al., 2001; Wang, 2003; Sousa et al., 2009; Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011) or with phylogenetic trees built by taking moments of the site frequency spectrum over large sets of SNPs (Reich et al., 2009; Green et al., 2010; Patterson et al., 2012; Pickrell and Pritchard, 2012; Lipson et al., 2012). For example, f-statistic-based three- and four-population tests for admixture (Reich et al., 2009; Green et al., 2010; Patterson et al., 2012) are highly sensitive in the proper parameter regimes and when the set of sampled populations sufficiently represents the phylogeny. One disadvantage of drift-based statistics, however, is that because the rate of genetic drift depends on population size, these methods do not allow for inference of the time that has elapsed since admixture events.

Finally, Moorjani et al. (2011) recently proposed a fourth approach, using associations between pairs of loci to make inference about admixture, which we further develop in this article. In general, linkage disequilibrium (LD) in a population can be generated by selection, genetic drift, or population structure, and it is eroded by recombination. Within a homogeneous population, steady-state neutral LD is maintained by the balance of drift and recombination, typically becoming negligible in humans at distances of more than a few hundred kilobases (Reich et al., 2001; The International HapMap Consortium, 2007). Even if a population is currently well-mixed, however, it can retain longer-range *admixture LD* (ALD) from admixture events in its history involving previously separated populations. ALD is caused by associations between nearby loci co-inherited on an intact chromosomal block from one of the ancestral mixing populations (Chakraborty and Weiss, 1988). Recombination breaks down these associations, leaving a signature of the time elapsed since admixture that can be probed by aggregating pairwise LD measurements through an appropriate weighting scheme; the resulting weighted LD curve (as a function of genetic distance) exhibits an exponential decay with rate constant giving the age of admixture (Moorjani et al., 2011; Patterson et al., 2012). This approach to admixture dating is similar in spirit to strategies based on local ancestry, but LD statistics have the advantage of a simple mathematical form that facilitates error analysis.

In this paper, we comprehensively develop LD-based admixture inference, extending the methodology to several novel applications that constitute a versatile set of tools for investigating admixture. We first propose a cleaner functional form of the underlying weighted LD statistic and provide a precise mathematical development of its properties. As an immediate result of this theory, we observe that our new weighted LD statistic can be used to infer mixture proportions as well as dates, extending the results of Pickrell et al. (2012). Moreover, such inference can still be performed (albeit with reduced power) when data are available from only the admixed population and one surrogate ancestral population, whereas all previous techniques require at least two such reference populations. As a second application, we present an LD-based three-population test for admixture with sensitivity complementary to the 3-population f-statistic test (Reich et al., 2009; Patterson et al., 2012) and characterize the scenarios in which each is advantageous. We further show that phylogenetic relationships between true mixing populations and present-day references can be inferred by comparing weighted LD curves using weights derived from a suite of reference populations. Finally, we describe several improvements to the computation and fitting of weighted LD curves: we show how to detect confounding LD from sources other than admixture, improving the robustness of our methods in the presence of such effects, and we present a novel fast Fourier transform-based algorithm for weighted LD computation that reduces typical run times from hours to seconds. We implement all of these advances in a software package, *ALDER* (Admixture-induced Linkage Disequilibrium for Evolutionary Relationships).

We demonstrate the performance of *ALDER* by using it to test for admixture among all HGDP populations (Li et al., 2008b) and compare its results to those of the 3-population test, highlighting the sensitivity trade-offs of each approach. We further illustrate our methodology with case studies of Central African Pygmies, Sardinians, and Japanese, revealing new details that add to our understanding of admixture events in the history of each population.

2.2 Methods

2.2.1 Properties of weighted admixture LD

In this section we introduce a weighted LD statistic that uses the decay of LD to detect signals of admixture given SNP data from an admixed population and reference populations. This statistic is similar to, but has an important difference from, the weighted LD statistic used in *ROLLOFF* (Moorjani et al., 2011; Patterson et al., 2012). The formulation of our statistic is particularly important in allowing us to use the amplitude (i.e., y-intercept) of the weighted LD curve to make inferences about history. We begin by deriving quantitative mathematical properties of this statistic that can be used to infer admixture parameters.

Basic model and notation

We will primarily consider a point-admixture model in which a population C' descends from a mixture of populations A and B to form C, n generations ago, in proportions $\alpha + \beta = 1$, followed by random mating (Figure 2.1). As we discuss later, we can assume for our purposes that the genetic drift between C and C' is negligible, and hence we will simply refer to the descendant population as C as well; we will state whether we mean the population

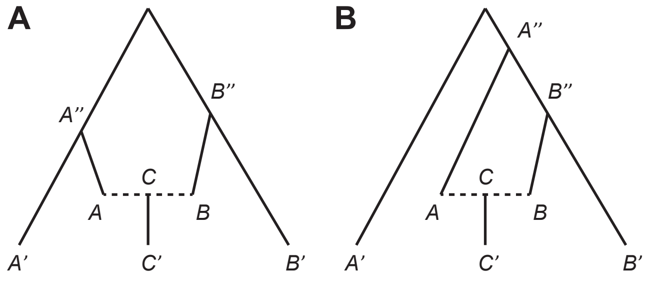


Figure 2.1. Notational diagram of phylogeny containing admixed population and references. Population C' is descended from an admixture between A and B to form C; populations A' and B' are present-day references. In practice, we assume that post-admixture drift is negligible, i.e., the C–C' branch is extremely short and C' and C have identical allele frequencies. The branch points of A' and B' from the A–B lineage are marked A'' and B''; note that in a rooted phylogeny, these need not be most recent common ancestors.

immediately after admixture vs. n generations later when there is any risk of ambiguity. We are interested in the properties of the LD in population C induced by admixture. Consider two biallelic, neutrally evolving SNPs x and y, and for each SNP call one allele '0' and the other '1' (this assignment is arbitrary; '0' and '1' do not need to be oriented with regard to ancestral state via an outgroup). Denote by $p_A(x)$, $p_B(x)$, $p_A(y)$, and $p_B(y)$ the frequencies of the '1' alleles at x and y in the mixing populations A and B (at the time of admixture), and let $\delta(x) := p_A(x) - p_B(x)$ and $\delta(y) := p_A(y) - p_B(y)$ be the allele frequency differences.

Let d denote the genetic distance between x and y and assume that x and y were in linkage equilibrium in populations A and B. Then the LD in population C immediately after admixture is

$$D_0 = \alpha\beta\delta(x)\delta(y),$$

where D is the standard haploid measure of linkage disequilibrium as the covariance of alleles at x and y (Chakraborty and Weiss, 1988). After n generations of random mating, the LD decays to

$$D_n = e^{-nd}D_0 = e^{-nd}\alpha\beta\delta(x)\delta(y)$$

assuming infinite population size (Chakraborty and Weiss, 1988). For a finite population, the above formula holds in expectation with respect to random drift, with a small adjustment factor caused by post-admixture drift (Ohta and Kimura, 1971):

$$E[D_n] = e^{-nd}e^{-n/2N_e}\alpha\beta\delta(x)\delta(y),$$

where N_e is the effective population size. In most applications the adjustment factor $e^{-n/2N_e}$ is negligible, so we will omit it in what follows (Moorjani et al., 2013, , Note S1).

In practice, our data consist of unphased diploid genotypes, so we expand our notation accordingly. Consider sampling a random individual from population C (n generations after admixture). We use a pair of $\{0, 1\}$ random variables X_1 and X_2 to refer to the two alleles at x and define random variables Y_1 and Y_2 likewise. Our unphased SNP data represent observations of the $\{0, 1, 2\}$ random variables $X := X_1 + X_2$ and $Y := Y_1 + Y_2$.

Define $z(x, y)$ to be the covariance

$$z(x, y) := \operatorname{cov}(X, Y) = \operatorname{cov}(X_1 + X_2, Y_1 + Y_2), \tag{2.1}$$

which can be decomposed into a sum of four haplotype covariances:

$$z(x, y) = \operatorname{cov}(X_1, Y_1) + \operatorname{cov}(X_2, Y_2) + \operatorname{cov}(X_1, Y_2) + \operatorname{cov}(X_2, Y_1). \tag{2.2}$$

The first two terms measure D for the separate chromosomes, while the third and fourth terms vanish, since they represent covariances between variables for different chromosomes, which are independent. Thus, the expectation (again with respect to random drift) of the total diploid covariance is simply

$$E[z(x, y)] = 2e^{-nd}\alpha\beta\delta(x)\delta(y). \tag{2.3}$$

Relating weighted LD to admixture parameters

Moorjani et al. (2011) first observed that pairwise LD measurements across a panel of SNPs can be combined to enable accurate inference of the age of admixture, n. The crux of their approach was to harness the fact that the ALD between two sites x and y scales as e^{-nd} multiplied by the product of allele frequency differences $\delta(x)\delta(y)$ in the mixing populations. While the allele frequency differences $\delta(\cdot)$ are usually not directly computable, they can often be approximated. Thus, Moorjani et al. (2011) formulated a method, *ROLLOFF*, that dates admixture by fitting an exponential decay e^{-nd} to correlation coefficients between LD measurements and surrogates for $\delta(x)\delta(y)$. Note that Moorjani et al. (2011) define $z(x, y)$ as a sample correlation coefficient, analogous to the classical LD measure r, as opposed to the sample covariance (2.1) we use here; we find the latter more mathematically convenient.

We build upon these previous results by deriving exact formulas for weighted sums of ALD under a variety of weighting schemes that serve as useful surrogates for $\delta(x)\delta(y)$ in practice. These calculations will allow us to interpret the magnitude of weighted ALD to obtain additional information about admixture parameters. Additionally, the theoretical development will generally elucidate the behavior of weighted ALD and its applicability in various phylogenetic scenarios.

Following Moorjani et al. (2011), we partition all pairs of SNPs (x, y) into bins of roughly constant genetic distance:

$$\mathcal{S}(d) := \left\{(x, y) : d - \frac{\epsilon}{2} < |x - y| < d + \frac{\epsilon}{2}\right\},$$

where ϵ is a discretization parameter inducing a discretization on d. Given a choice of weights $w(\cdot)$, one per SNP, we define the weighted LD at distance d as

$$a(d) := \frac{\sum_{\mathcal{S}(d)} z(x, y)w(x)w(y)}{|\mathcal{S}(d)|}.$$

Assume first that our weights are the true allele frequency differences in the mixing populations, i.e., $w(x) = \delta(x)$ for all x. Applying (2.3),

$$\begin{aligned} E[a(d)] &= E\left[\frac{\sum_{\mathcal{S}(d)} z(x,y)\delta(x)\delta(y)}{|\mathcal{S}(d)|}\right] \\ &= \frac{\sum_{\mathcal{S}(d)} 2\alpha\beta E[\delta(x)^2\delta(y)^2]e^{-nd}}{|\mathcal{S}(d)|} \\ &= 2\alpha\beta F_2(A,B)^2 e^{-nd}, \end{aligned} \tag{2.4}$$

where $F_2(A, B)$ is the expected squared allele frequency difference for a randomly drifting neutral allele, as defined in Reich et al. (2009) and Patterson et al. (2012). Thus, $a(d)$ has the form of an exponential decay as a function of d, with time constant n giving the date of admixture.

In practice, we must compute an empirical estimator of $a(d)$ from a finite number of sampled genotypes. Say we have a set of m diploid admixed samples from population C indexed by $i = 1, \ldots, m$, and denote their genotypes at sites x and y by $x_i, y_i \in \{0, 1, 2\}$. Also assume we have some finite number of reference individuals from A and B with empirical mean allele frequences $\hat{p}_A(\cdot)$ and $\hat{p}_B(\cdot)$. Then our estimator is

$$\hat{a}(d) := \frac{\sum_{\mathcal{S}(d)} \widehat{\operatorname{cov}(X,Y)}(\hat{p}_A(x) - \hat{p}_B(x))(\hat{p}_A(y) - \hat{p}_B(y))}{|\mathcal{S}(d)|}, \tag{2.5}$$

where

$$\widehat{\operatorname{cov}(X,Y)} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})$$

is the usual unbiased sample covariance, so the expectation over the choice of samples satisfies $E[\hat{a}(d)] = a(d)$ (assuming no background LD, so the ALD in population C is independent of the drift processes producing the weights).

The weighted sum $\sum_{\mathcal{S}(d)} z(x, y)w(x)w(y)$ is a natural quantity to use for detecting ALD decay and is common to our weighted LD statistic $\hat{a}(d)$ and previous formulations of the *ROLLOFF* statistic. Indeed, for SNP pairs (x, y) at a fixed distance d, we can think of equation (2.3) as providing a simple linear regression model between LD measurements $z(x, y)$ and allele frequency divergence products $\delta(x)\delta(y)$. In practice, the linear relation is made noisy by random sampling, as noted above, but the regression coefficient $2\alpha\beta e^{-nd}$ can be inferred by combining measurements from many SNP pairs (x, y). In fact, the weighted sum $\sum_{\mathcal{S}(d)} \hat{z}(x, y)\hat{\delta}(x)\hat{\delta}(y)$ in the numerator of formula (2.5) is precisely the numerator of the least-squares estimator of the regression coefficient, which is the formulation of *ROLLOFF* given in Moorjani et al. (2013, , Note S1). Note that measurements of $z(x, y)$ cannot be combined directly without a weighting scheme, as the sign of the LD can be either positive or negative; additionally, the weights tend to preserve signal from ALD while depleting contributions from other forms of LD.

Up to scaling, our *ALDER* formulation is roughly equivalent to the regression coefficient formulation of *ROLLOFF* (Moorjani et al., 2013, , Note S1). In contrast, the original *ROLLOFF* statistic (Patterson et al., 2012) computed a *correlation* coefficient between

$z(x, y)$ and $w(x)w(y)$ over $\mathcal{S}(d)$. However, the normalization term $\sqrt{\sum_{\mathcal{S}(d)} z(x, y)^2}$ in the denominator of the correlation coefficient can exhibit an unwanted d-dependence that biases the inferred admixture date if the admixed population has undergone a strong bottleneck (Moorjani et al., 2013, , Note S1) or in the case of recent admixture and large sample sizes. Beyond correcting the date bias, the $\hat{a}(d)$ curve that *ALDER* computes has the advantage of a simple form for its amplitude in terms of meaningful quantities, providing us additional leverage on admixture parameters. Additionally, we will show that $\hat{a}(d)$ can be computed efficiently via a new fast Fourier transform-based algorithm.

Using weights derived from diverged reference populations

In the above development, we set the weights $w(x)$ to equal the allele frequency differences $\delta(x)$ between the true mixing populations A and B. In practice, in the absence of DNA samples from past populations, it is impossible to measure historical allele frequencies from the time of mixture, so instead, we substitute reference populations A' and B' that are accessible, setting $w(x) = \delta'(x) := p_{A'}(x) - p_{B'}(x)$. In a given data set, the closest surrogates A' and B' may be somewhat diverged from A and B, so it is important to understand the consequences for the weighted LD $a(d)$.

We show in Appendix B.1 that with reference populations A' and B' in place of A and B, equation (2.4) for the expected weighted LD curve changes only slightly, becoming

$$E[a(d)] = 2\alpha\beta F_2(A'', B'')^2 e^{-nd}, \tag{2.6}$$

where A'' and B'' are the branch points of A' and B' on the A–B lineage (Figure 2.1). Notably, the curve still has the form of an exponential decay with time constant n (the age of admixture), albeit with its amplitude (and therefore signal-to-noise ratio) attenuated according to how far A'' and B'' are from the true ancestral mixing populations. Drift along the A'–A'' and B'–B'' branches likewise decreases signal-to-noise but in the reverse manner: higher drift on these branches makes the weighted LD curve noisier but does not change its expected amplitude (Figure 2.2; see Appendix B.3 for additional discussion). As above, given a real data set containing finite samples, we compute an estimator $\hat{a}(d)$ analogous to formula (2.5) that has the same expectation (over sampling and drift) as the expectation of $a(d)$ with respect to drift (2.6).

Using the admixed population as one reference

Weighted LD can also be computed with only a single reference population by using the admixed population as the other reference (Pickrell et al., 2012, , Supplement Sec. 4). Assuming first that we know the allele frequencies of the ancestral mixing population A and the admixed population C, the formula for the expected curve becomes

$$E[a(d)] = 2\alpha\beta^3 F_2(A, B)^2 e^{-nd}. \tag{2.7}$$

Using C itself as one reference population and R' as the other reference (which could branch anywhere between A and B), the formula for the amplitude is slightly more complicated,

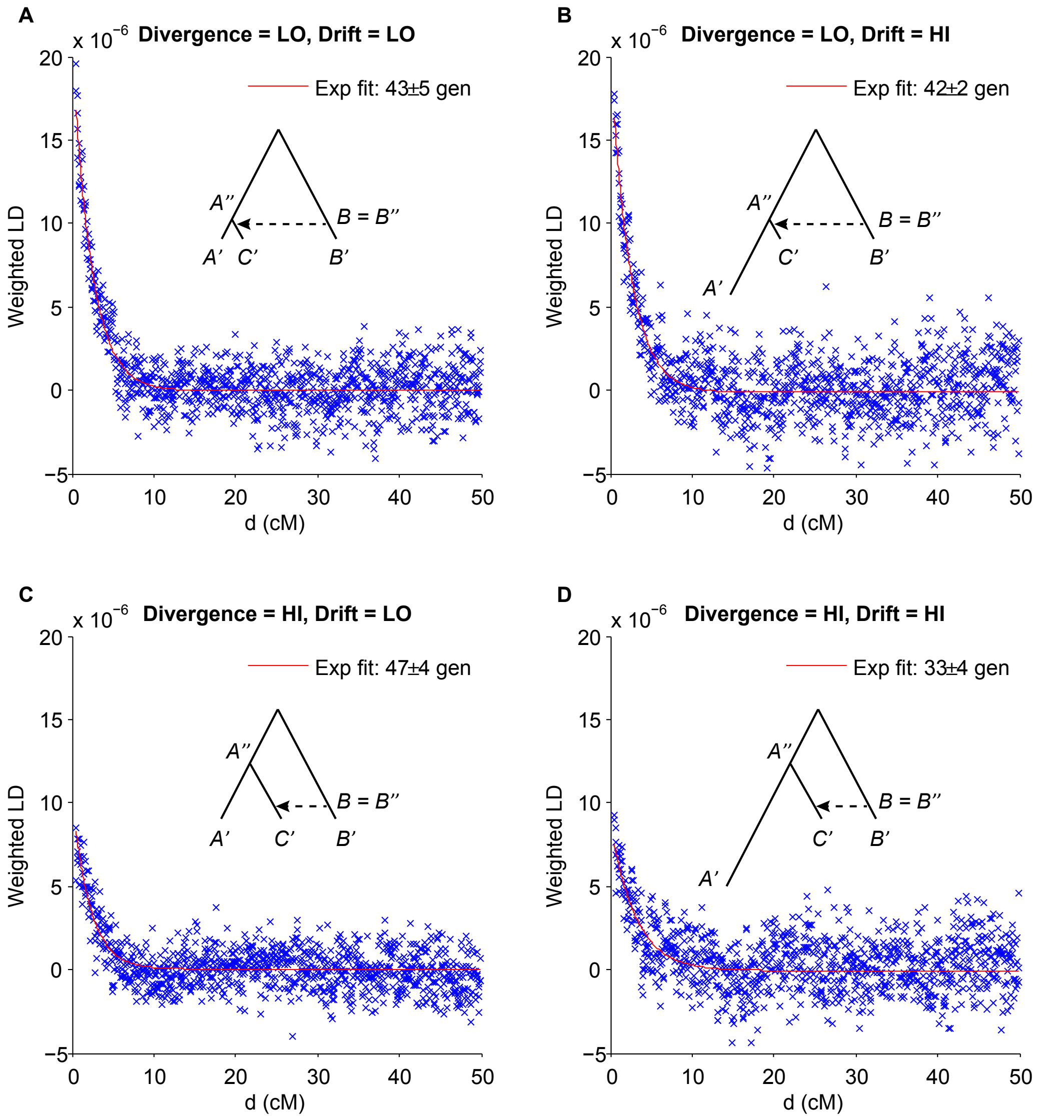


Figure 2.2. Weighted LD curves from four coalescent simulations of admixture scenarios with varying divergence times and drift between the reference population A' and the true mixing population. In each case, gene flow occurred 40 generations ago. In the low-divergence scenarios, the split point A'' is immediately prior to gene flow, while in the high-divergence scenarios, A'' is halfway up the tree (520 generations ago). The high-drift scenarios are distinguished from the low-drift scenarios by a 20-fold reduction in population size for the past 40 generations. Standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation.

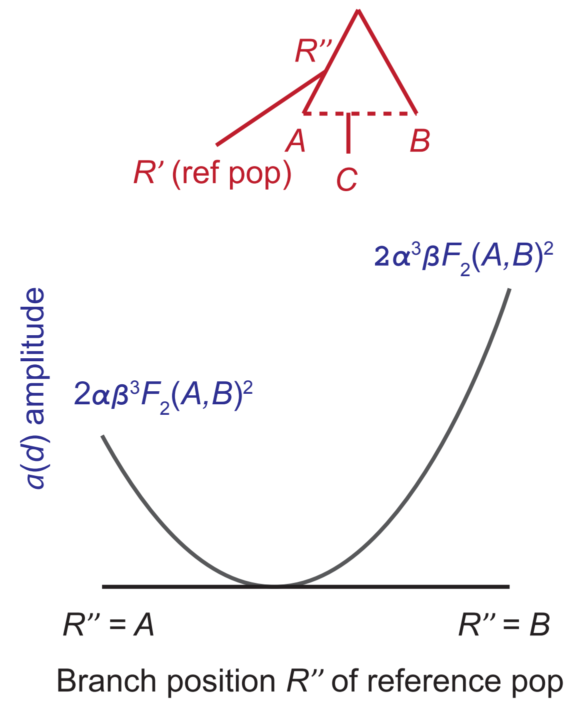

R″
A
B
C
R′ (ref pop)
$2\alpha^3\beta F_2(A,B)^2$
$2\alpha\beta^3F_2(A,B)^2$
a(d) amplitude
R″ = A
R″ = B
Branch position R″ of reference pop

Figure 2.3. Dependence of single-reference weighted LD amplitude on the reference population. When taking weights as allele frequency differences between the admixed population and a single reference population R', the weighted LD curve $a(d)$ has expected amplitude proportional to $(\alpha F_2(A, R'') - \beta F_2(B, R''))^2$, where R'' is the point along the A–B lineage at which the reference population branches. Note in particular that as R'' varies from A to B, the amplitude traces out a parabola that starts at $2\alpha\beta^3 F_2(A, B)^2$, decreases to a minimum value of 0, and increases to $2\alpha^3\beta F_2(A, B)^2$.

but the curve retains the e^{-nd} decay (Figure 2.3):

$$E[a(d)] = 2\alpha\beta(\alpha F_2(A, R'') - \beta F_2(B, R''))^2 e^{-nd}. \tag{2.8}$$

Derivations of these formulas are given in Appendix B.1.

A subtle but important technical issue arises when computing weighted LD with a single reference. In this case, the true weighted LD statistic is

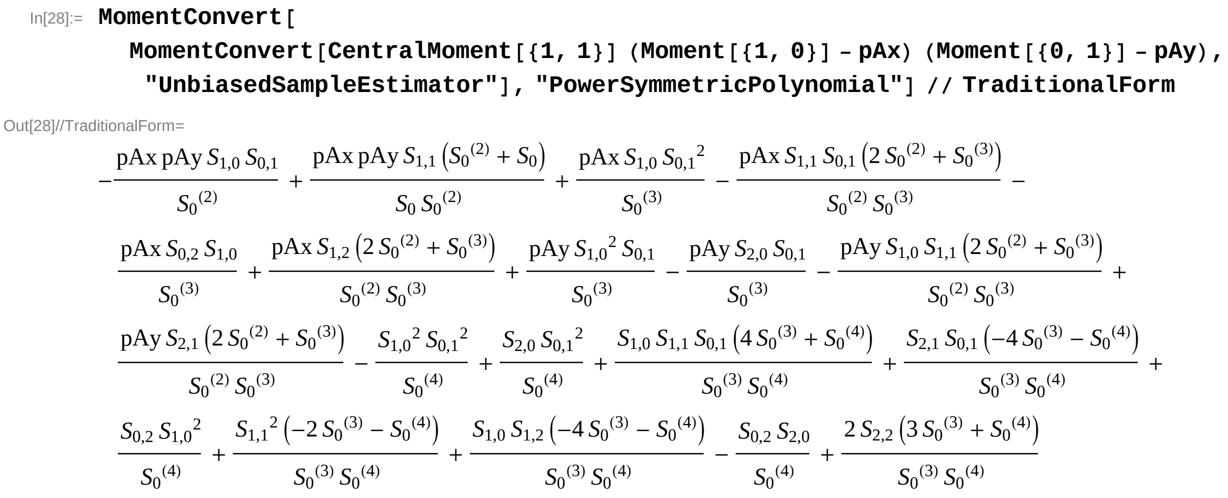
$$a(d) = \mathrm{cov}(X, Y)(\mu_x - p(x))(\mu_y - p(y)),$$

where

$$\mu_x = \alpha p_A(x) + \beta p_B(x) \quad \text{and} \quad \mu_y = \alpha p_A(y) + \beta p_B(y)$$

are the mean allele frequencies of the admixed population (ignoring drift) and $p(\cdot)$ denotes allele frequencies of the reference population. Here $a(d)$ cannot be estimated accurately by the naïve formula

$$\widehat{\mathrm{cov}(X, Y)}(\hat{\mu}_x - \hat{p}(x))(\hat{\mu}_y - \hat{p}(y)),$$



```
In[28]:= MomentConvert[
           MomentConvert[CentralMoment[{1, 1}] (Moment[{1, 0}] - pAx) (Moment[{0, 1}] - pAy),
            "UnbiasedSampleEstimator"], "PowerSymmetricPolynomial"] // TraditionalForm
```

Out[28]//TraditionalForm=

$$-\frac{\mathrm{pAx}\,\mathrm{pAy}\,S_{1,0}\,S_{0,1}}{S_0{}^{(2)}}+\frac{\mathrm{pAx}\,\mathrm{pAy}\,S_{1,1}\left(S_0{}^{(2)}+S_0\right)}{S_0\,S_0{}^{(2)}}+\frac{\mathrm{pAx}\,S_{1,0}\,S_{0,1}{}^2}{S_0{}^{(3)}}-\frac{\mathrm{pAx}\,S_{1,1}\,S_{0,1}\left(2\,S_0{}^{(2)}+S_0{}^{(3)}\right)}{S_0{}^{(2)}\,S_0{}^{(3)}}-$$
$$\frac{\mathrm{pAx}\,S_{0,2}\,S_{1,0}}{S_0{}^{(3)}}+\frac{\mathrm{pAx}\,S_{1,2}\left(2\,S_0{}^{(2)}+S_0{}^{(3)}\right)}{S_0{}^{(2)}\,S_0{}^{(3)}}+\frac{\mathrm{pAy}\,S_{1,0}{}^2\,S_{0,1}}{S_0{}^{(3)}}-\frac{\mathrm{pAy}\,S_{2,0}\,S_{0,1}}{S_0{}^{(3)}}-\frac{\mathrm{pAy}\,S_{1,0}\,S_{1,1}\left(2\,S_0{}^{(2)}+S_0{}^{(3)}\right)}{S_0{}^{(2)}\,S_0{}^{(3)}}+$$
$$\frac{\mathrm{pAy}\,S_{2,1}\left(2\,S_0{}^{(2)}+S_0{}^{(3)}\right)}{S_0{}^{(2)}\,S_0{}^{(3)}}-\frac{S_{1,0}{}^2\,S_{0,1}{}^2}{S_0{}^{(4)}}+\frac{S_{2,0}\,S_{0,1}{}^2}{S_0{}^{(4)}}+\frac{S_{1,0}\,S_{1,1}\,S_{0,1}\left(4\,S_0{}^{(3)}+S_0{}^{(4)}\right)}{S_0{}^{(3)}\,S_0{}^{(4)}}+\frac{S_{2,1}\,S_{0,1}\left(-4\,S_0{}^{(3)}-S_0{}^{(4)}\right)}{S_0{}^{(3)}\,S_0{}^{(4)}}+$$
$$\frac{S_{0,2}\,S_{1,0}{}^2}{S_0{}^{(4)}}+\frac{S_{1,1}{}^2\left(-2\,S_0{}^{(3)}-S_0{}^{(4)}\right)}{S_0{}^{(3)}\,S_0{}^{(4)}}+\frac{S_{1,0}\,S_{1,2}\left(-4\,S_0{}^{(3)}-S_0{}^{(4)}\right)}{S_0{}^{(3)}\,S_0{}^{(4)}}-\frac{S_{0,2}\,S_{2,0}}{S_0{}^{(4)}}+\frac{2\,S_{2,2}\left(3\,S_0{}^{(3)}+S_0{}^{(4)}\right)}{S_0{}^{(3)}\,S_0{}^{(4)}}$$

Figure 2.4. Unbiased polyache estimator for weighted LD using the admixed population itself as one reference. Mathematica code and output are shown for computing the polyache statistic that estimates the one-reference weighted LD, $E[(X-\mu_x)(Y-\mu_y)(\mu_x-p_A(x))(\mu_y-p_A(y))]$, where $p_A(\cdot)$ are allele frequencies of the single reference population and μ_x and μ_y denote allele frequencies of the admixed population. In the above, $S_0^{(k)} := m(m-1)\cdots(m-k+1)$ and $S_{r,s} := \sum_{i=1}^m X_i^r Y_i^s$, where m is the number of admixed samples and i ranges over the admixed individuals, which have allele counts X_i and Y_i at sites x and y.

which is the natural analog of (2.5). The difficulty is that the covariance term and the weights both involve the allele frequencies μ_x and μ_y; thus, while the standard estimators for each term are individually unbiased, their product is a biased estimate of the weighted LD.

Pickrell et al. (2012) circumvents this problem by partitioning the admixed samples into two groups, designating one group for use as admixed representatives and the other as a reference population; this method eliminates bias but reduces statistical power. We instead compute a polyache statistic (Figure 2.4) that provides an unbiased estimator $\hat{a}(d)$ of the weighted LD with maximal power.

Affine term in weighted LD curve from population substructure

Weighted LD curves computed on real populations often exhibit a nonzero horizontal asymptote contrary to the exact exponential decay formulas we have derived above. Such behavior can be caused by assortative mating resulting in subpopulations structured by ancestry percentage in violation of our model. We show in Appendix B.1 that if we instead model the admixed population as consisting of randomly mating subpopulations with heterogeneous amounts α—now a random variable—of mixed ancestry, our equations for the curves take the form

$$E[a(d)] = Me^{-nd} + K, \tag{2.9}$$

where M is a coefficient representing the contribution of admixture LD and K is an additional constant produced by substructure. Conveniently, however, the sum $M + K/2$ satisfies the same equations that the coefficient of the exponential does in the homogeneous case: adjusting equation (2.6) for population substructure gives

$$M + K/2 = 2\alpha\beta F_2(A'', B'')^2 \tag{2.10}$$

for two-reference weighted LD, and in the one-reference case, modifying equation (2.8) gives

$$M + K/2 = 2\alpha\beta(\alpha F_2(A, R'') - \beta F_2(B, R''))^2. \tag{2.11}$$

For brevity, from here on we will take the amplitude of an exponential-plus-affine curve to mean $M + K/2$.

2.2.2 Admixture inference using weighted LD

We now describe how the theory we have developed can be used to investigate admixture. We detail novel techniques that use weighted LD to infer admixture parameters, test for admixture, and learn about phylogeny.

Inferring admixture dates and fractions using one or two reference populations

As noted above, our *ALDER* formulation of weighted LD hones the original two-reference admixture dating technique of *ROLLOFF* (Moorjani et al., 2011), correcting a possible bias (Moorjani et al., 2013, , Note S1), and the one-reference technique (Pickrell et al., 2012), improving statistical power. Pickrell et al. (2012) also observed that weighted LD can be used to estimate ancestral mixing fractions. We further develop this application now.

The main idea is to treat our expressions for the amplitude of the weighted LD curve as equations that can be solved for the ancestry fractions α and $\beta = 1 - \alpha$. First consider two-reference weighted LD. Given samples from an admixed population C and reference populations A' and B', we compute the curve $\hat{a}(d)$ and fit it as an exponential decay plus affine term: $\hat{a}(d) \approx \hat{M}e^{-nd} + \hat{K}$. Let $\hat{a}_0 := \hat{M} + \hat{K}/2$ denote the amplitude of the curve. Then equation (2.10) gives us a quadratic equation that we can solve to obtain an estimate $\hat{\alpha}$ of the mixture fraction α,

$$2\hat{\alpha}(1 - \hat{\alpha})F_2(A'', B'')^2 = \hat{a}_0,$$

assuming we can estimate $F_2(A'', B'')^2$. Typically the branch-point populations A'' and B'' are unavailable, but their F_2 distance can be computed by means of an admixture tree (Patterson et al., 2012; Pickrell and Pritchard, 2012; Lipson et al., 2012). A caveat of this approach is that α and $1 - \alpha$ produce the same amplitude and cannot be distinguished by this method alone; additionally, the inversion problem is ill-conditioned near $\alpha = 0.5$, where the derivative of the quadratic vanishes.

The situation is more complicated when using the admixed population as one reference.

First, the amplitude relation from equation (2.11) gives a quartic equation in $\hat{\alpha}$:

$$2\hat{\alpha}(1-\hat{\alpha})[\hat{\alpha}F_2(A,R'')-(1-\hat{\alpha})F_2(B,R'')]^2=\hat{a}_0.$$

Second, the F_2 distances involved are in general not possible to calculate by solving allele frequency moment equations (Patterson et al., 2012; Lipson et al., 2012). In the special case that one of the true mixing populations is available as a reference, however—i.e., $R'=A$—Pickrell et al. (2012) demonstrated that mixture fractions can be estimated much more easily. From equation (2.7), the expected amplitude of the curve is $2\alpha\beta^3F_2(A,B)^2$. On the other hand, assuming no drift in C since the admixture, allele frequencies in C are given by weighted averages of allele frequencies in A and B with weights α and β; thus, the squared allele frequency differences from A to B and C satisfy

$$F_2(A,C)=\beta^2F_2(A,B),$$

and $F_2(A,C)$ is estimable directly from the sample data. Combining these relations, we can obtain our estimate $\hat{\alpha}$ by solving the equation

$$2\hat{\alpha}/(1-\hat{\alpha})=\hat{a}_0/F_2(A,C)^2. \tag{2.12}$$

In practice, the true mixing population A is not available for sampling, but a closely-related population A' may be. In this case, the value of $\hat{\alpha}$ given by equation (2.12) with A' in place of A is a lower bound on the true mixture fraction α (see Appendix B.1 for theoretical development and Results for simulations exploring the tightness of the bound). This bounding technique is the most compelling of the above mixture fraction inference approaches, as prior methods cannot perform such inference with only one reference population. In contrast, when more references are available, moment-based admixture tree-fitting methods, for example, readily estimate mixture fractions (Patterson et al., 2012; Pickrell and Pritchard, 2012; Lipson et al., 2012). In such cases we believe that existing methods are more robust than LD-based inference, which suffers from the degeneracy of solutions noted above; however, the weighted LD approach can provide confirmation based on a different genetic mechanism.

Testing for admixture

Thus far, we have taken it as given that the population C of interest is admixed and developed methods for inferring admixture parameters by fitting weighted LD curves. Now we consider the question of whether weighted LD can be used to determine whether admixture occurred in the first place. We develop a weighted LD-based formal test for admixture that is broadly analogous to the drift-based 3-population test (Reich et al., 2009; Patterson et al., 2012) but sensitive in different scenarios.

A complication of interpreting weighted LD is that certain demographic events other than admixture can also produce positive weighted LD that decays with genetic distance, particularly in the one-reference case. Specifically, if population C has experienced a recent bottleneck or an extended period of low population size, it may contain long-range LD. Furthermore, this LD typically has some correlation with allele frequencies in C; consequently, using C as one reference in the weighting scheme produces a spurious weighted LD signal.

In the two-reference case, LD from reduced population size in C is generally washed out by the weighting scheme assuming the reference populations A' and B' are not too genetically similar to C. The reason is that the weights $\delta(\cdot) = p_{A'}(\cdot) - p_{B'}(\cdot)$ arise from drift between A' and B' that is independent of demographic events producing LD in C (beyond genetic distances that are so short that the populations share haplotypes descended without recombination from their common ancestral haplotype). Thus, observing a two-reference weighted LD decay curve is generally good evidence that population C is admixed. There is still a caveat, however. If C and one of the references, say A', share a recent population bottleneck, then the bottleneck-induced LD in C can be correlated to the allele frequencies of A', resulting once again in spurious weighted LD. In fact, the one-reference example mentioned above is the limiting case $A' = C$ of this situation.

With these considerations in mind, we propose an LD-based three-population test for admixture that includes a series of pre-tests safeguarding against the pathological demographies that can produce a non-admixture weighted LD signal. We outline the test now; details are in Appendix B.2. Given a population C to test for admixture and references A' and B', the main test is whether the observed weighted LD $\hat{a}(d)$ using A'–B' weights is positive and well-fit by an exponential decay curve. We estimate a jackknife-based p-value by leaving out each chromosome in turn and re-fitting the weighted LD as an exponential decay; the jackknife then gives us a standard error on the fitting parameters—namely, the amplitude and the decay constant—that we use to measure the significance of the observed curve.

The above procedure allows us to determine whether there is sufficient signal in the weighted LD curve to reject the null hypothesis (under which $\hat{a}(d)$ is random "colored" noise in the sense that it contains autocorrelation). However, in order to conclude that the curve is the result of admixture, we must rule out the possibility that it is being produced by demography unrelated to admixture. We therefore apply the following pre-test procedure. First, we determine the distance to which LD in C is significantly correlated with LD in either A' or B'; to minimize signal from shared demography, we subsequently ignore data from SNP pairs at distances smaller than this correlation threshold. Then, we compute one-reference weighted LD curves for population C with A'–C and B'–C weights and check that the curves are well-fit as exponential decays. In the case that C is actually admixed between populations related to A' and B', the one-reference A'–C and B'–C curves pick up the same e^{-nd} admixture LD decay signal. If C is not admixed but has experienced a shared bottleneck with A' (producing false-positive admixture signals from the A'–B' and B'–C curves), however, the A'–C weighting scheme is unlikely to produce a weighted LD curve, especially when fitting beyond the LD correlation threshold.

This test procedure is intended to be conservative, so that a population C identified as admixed can strongly be assumed to be so, whereas if C is not identified as admixed, we are less confident in claiming that C has experienced no admixture whatsoever. In situations where distinguishing admixture from other demography is particularly difficult, the test will err on the side of caution; for example, even if C is admixed, the test may fail to identify C as admixed if it has also experienced a bottleneck. Also, if a reference A' shares some of the same admixture history as C or is simply very closely related to C, the pre-test will typically identify long-range correlated LD and deem A' an unsuitable reference to use for testing admixture. The behavior of the test and pre-test criteria are explored in detail with coalescent simulations in Appendix B.3.

Learning about phylogeny

Given a triple of populations $(C; A', B')$, our test can identify admixture in the test population C, but what does this imply about the relationship of populations A' and B' to C? As with the drift-based 3-population test, test results must be interpreted carefully: even if C is admixed, this does not necessarily mean that the reference populations A' and B' are closely related to the true mixing populations. However, computing weighted LD curves with a suite of different references can elucidate the phylogeny of the populations involved, since our amplitude formulas (2.10) and (2.11) provide information about the locations on the phylogeny at which the references diverge from the true mixing populations.

More precisely, in the notation of Figure 2.1, the amplitude of the two-reference weighted LD curve is $2\alpha\beta F_2(A'', B'')^2$, which is maximized when $A'' = A$ and $B'' = B$ and is minimized when $A'' = B''$. So, for example, we can fix A' and compute curves for a variety of references B'; the larger the resulting amplitude, the closer the branch point B'' is to B. In the one-reference case, as the reference R' is varied, the amplitude $2\alpha\beta(\alpha F_2(A, R'') - \beta F_2(B, R''))^2$ traces out a parabola that starts at $2\alpha\beta^3 F_2(A, B)^2$ when $R'' = A$, decreases to a minimum value of 0, and increases again to $2\alpha^3\beta F_2(A, B)^2$ when $R'' = B$ (Figure 2.3). Here, the procedure is more qualitative because the branches $F_2(A, R'')$ and $F_2(B, R'')$ are less directly useful and the mixture proportions α and β may not be known.

2.2.3 Implementation of *ALDER*

We now describe some more technical details of the *ALDER* software package in which we have implemented our weighted LD methods.

Fast Fourier transform algorithm for computing weighted LD

We developed a novel algorithm that algebraically manipulates the weighted LD statistic into a form that can be computed using a fast Fourier transform (FFT), dramatically speeding up the computation (File B.4). The algebraic transformation is made possible by the simple form (2.5) of our weighted LD statistic along with a genetic distance discretization procedure that is similar in spirit to *ROLLOFF* (Moorjani et al., 2011) but subtly different: instead of binning the contributions of SNP pairs (x, y) by discretizing the genetic distance $|x-y| = d$, we discretize the genetic map positions x and y themselves (using a default resolution of 0.05 cM) (Figure 2.5). For two-reference weighted LD, the resulting FFT-based algorithm that we implemented in *ALDER* has computational cost that is approximately linear in the data size; in practice, it ran three orders of magnitude faster than *ROLLOFF* on typical data sets we analyzed.

Curve-fitting

We fit discretized weighted LD curves $\hat{a}(d)$ as $\hat{M}e^{-nd} + \hat{K}$ from equation (2.9), using least-squares to find best-fit parameters. This procedure is similar to *ROLLOFF*, but *ALDER* makes two important technical advances that significantly improve the robustness of the

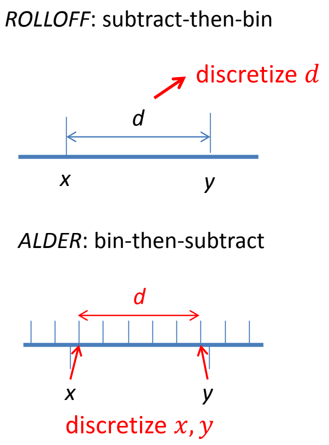


Figure 2.5. Comparison of binning procedures used by *ROLLOFF* and *ALDER*. Instead of discretizing inter-SNP distances, *ALDER* discretizes the genetic map before subtracting SNP coordinates.

fitting. First, *ALDER* directly estimates the affine term K that arises from the presence of subpopulations with differing ancestry percentages by using inter-chromosome SNP pairs that are effectively at infinite genetic distance (Appendix B.1). The algorithmic advances we implement in *ALDER* enable efficient computation of the average weighted LD over all pairs of SNPs on different chromosomes, giving $\hat{K}$ and, importantly, eliminating one parameter from the exponential fitting. In practice, we have observed that *ROLLOFF* fits are sometimes sensitive to the maximum inter-SNP distance d to which the weighted LD curve is computed and fit; *ALDER* eliminates this sensitivity.

Second, because background LD is present in real populations at short genetic distances and confounds the ALD signal (interfering with parameter estimates or producing spurious signal entirely), it is important to fit weighted LD curves starting only at a distance beyond which background LD is negligible. *ROLLOFF* used a fixed threshold of $d > 0.5$ cM, but some populations have longer-range background LD (e.g., from bottlenecks), and moreover, if a reference population is closely related to the test population, it can produce a spurious weighted LD signal due to recent shared demography. *ALDER* therefore estimates the extent to which the test population shares correlated LD with the reference(s) and only fits the weighted LD curve beyond this minimum distance as in our test for admixture (Appendix B.2).

We estimate standard errors on parameter estimates by performing a jackknife over the autosomes used in the analysis, leaving out each in turn. Note that the weighted LD measurements from individual pairs of SNPs that go into the computed curve $\hat{a}(d)$ are not independent of each other; however, the contributions of different chromosomes can reasonably be assumed to be independent.

2.2.4 Data sets

We primarily applied our weighted LD techniques to a data set of 940 individuals in 53 populations from the CEPH-Human Genome Diversity Cell Line Panel (HGDP) (Rosenberg et al., 2002) genotyped on an Illumina 650K SNP array (Li et al., 2008b). To study the effect of SNP ascertainment, we also analyzed the same HGDP populations genotyped on the Affymetrix Human Origins Array (Patterson et al., 2012). For some analyses we also included HapMap Phase 3 data (The International HapMap Consortium, 2010) merged either with the Illumina HGDP data set, leaving approximately 600K SNPs, or with the Indian data set of Reich et al. (2009) including 16 Andaman Islanders (9 Onge and 7 Great Andamanese), leaving 500K SNPs.

We also constructed simulated admixed chromosomes from 112 CEU and 113 YRI phased HapMap individuals using the following procedure, described in Moorjani et al. (2011). Given desired ancestry proportions α and β, the age n of the point admixture, and the number m of admixed individuals to simulate, we built each admixed chromosome as a composite of chromosomal segments from the source populations, choosing breakpoints via a Poisson process with rate constant n, and sampling blocks at random according to the specified mixture fractions. We stipulated that no individual haplotype could be reused at a given locus among the m simulated individuals, preventing unnaturally long identical-by-descent segments but effectively eliminating post-admixture genetic drift. For the short time scales we study (admixture occurring 200 or fewer generations ago), this approximation has little impact. We used this method in order to maintain some of the complications inherent in real data.

2.3 Results

2.3.1 Simulations

First, we demonstrate the accuracy of several forms of inference from *ALDER* on simulated data. We generated simulated genomes for mixture fractions of 75% YRI / 25% CEU and 90% YRI / 10% CEU and admixture dates of 10, 20, 50, 100, and 200 generations ago. For each mixture scenario we simulated 40 admixed individuals according to the procedure above.

We first investigated the admixture dates estimated by *ALDER* using a variety of reference populations drawn from the HGDP with varying levels of divergence from the true mixing populations. On the African side, we used HGDP Yoruba (21 samples; essentially the same population as HapMap YRI) and San (5 samples); on the European side, we used French (28 samples; very close to CEU), Han (34 samples), and Papuan (17 samples). We computed two-reference weighted LD curves using pairs of references, one from each group, as well as one-reference curves using the simulated population as one reference and each of the above HGDP populations as the other.

For the 75% YRI mixture, estimated dates are nearly all accurate to within 10% (Table 2.1). The noise levels of the fitted dates (estimated by *ALDER* using the jackknife) are the lowest for the Yoruba–French curve, as expected, followed by the one-reference curve with French, consistent with the admixed population being mostly Yoruba. The situation is

Table 2.1. Dates of admixture estimated for simulated 75% YRI / 25% CEU mixtures.

Ref 1	Ref 2	10	20	50	100	200
Yoruba	French	9±1	20±1	49±2	107±5	195±9
Yoruba	Han	9±1	21±1	50±2	107±6	191±12
Yoruba	Papuan	9±1	21±1	49±3	118±8	223±23
San	French	9±1	20±1	50±2	109±4	197±15
San	Han	9±0	21±1	51±3	111±4	194±16
San	Papuan	9±1	21±1	51±3	115±6	209±16
Yoruba		9±1	21±1	48±2	107±5	181±17
San		9±1	20±2	56±7	139±22	213±97
French		9±1	20±1	50±2	108±3	194±9
Han		9±0	21±1	52±2	110±6	192±17
Papuan		9±1	21±1	53±3	125±8	217±26

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

similar but noisier for the 90% YRI mixture (Table 2.2); in this case, the one-reference signal is quite weak with Yoruba and undetectable with San as the reference, due to the scaling of the amplitude (equation (2.11)) with the cube of the CEU mixture fraction.

We also compared fitted amplitudes of the weighted LD curves for the same scenarios to those predicted by formulas (2.10) and (2.11); the accuracy trends are similar (Tables 2.3, 2.4).

Finally, we tested formula (2.12) for bounding mixture proportions using one-reference weighted LD amplitudes. We computed lower bounds on the European ancestry fraction using French, Russian, Sardinian, and Kalash as successively more diverged references. As expected, the bounds are tight for the French reference and grow successively weaker (Tables 2.5, 2.6).

We also tried lower-bounding the African ancestry using one-reference curves with an African reference. In general, we expect lower bounds computed for the major ancestry proportion to be much weaker (Appendix B.1), and indeed we find this to be the case, with the only slightly diverged Mandenka population producing extremely weak bounds. An added complication is that the Mandenka are an admixed population with a small amount of West Eurasian ancestry (Price et al., 2009), which is not accounted for in the amplitude formulas we use here.

Another notable feature of *ALDER* is that, to a much greater extent than f-statistic methods, its inference quality improves with more samples from the admixed test population. As a demonstration of this, we simulated a larger set of 100 admixed individuals as above, for both 75% YRI / 25% CEU and 90% YRI / 10% CEU scenarios, and compared the date estimates obtained on subsets of 5–100 of these individuals with two different reference pairs (Tables 2.7, 2.8). With larger sample sizes, the estimates become almost uniformly

Table 2.2. Dates of admixture estimated for simulated 90% YRI / 10% CEU mixtures.

Ref 1	Ref 2	10	20	50	100	200
Yoruba	French	10±0	21±1	50±2	107±7	193±19
Yoruba	Han	10±0	20±1	51±2	109±10	220±32
Yoruba	Papuan	10±0	22±1	53±3	111±11	233±65
San	French	10±0	21±1	51±2	112±6	223±19
San	Han	10±0	21±1	52±3	121±5	254±40
San	Papuan	11±0	23±1	53±3	126±8	287±56
Yoruba		9±1	20±2	55±7	100±27	363±183
San		98±87	56±28	94±69	2±0	9±5
French		10±0	21±1	51±2	107±5	217±13
Han		11±0	21±1	52±2	111±7	234±25
Papuan		11±0	23±1	56±3	117±8	256±47

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table 2.3. Amplitudes of weighted LD curves (multiplied by 10^6) for simulated 75% YRI / 25% CEU mixtures.

Ref 1	Ref 2	Expected	10 gen	20 gen	50 gen	100 gen	200 gen
Yoruba	French	1173	1139±20	1203±40	1188±54	1283±100	1202±88
Yoruba	Han	693	678±17	717±28	711±43	774±73	716±74
Yoruba	Papuan	602	598±13	631±23	595±34	775±96	835±152
San	French	1017	981±23	1028±34	1044±49	1128±70	1037±130
San	Han	574	556±18	590±24	604±42	667±39	626±65
San	Papuan	491	487±17	514±20	503±34	589±45	574±60
Yoruba		75	77±2	81±4	74±4	83±6	71±13
San		40	40±3	42±3	50±6	66±13	43±34
French		655	626±12	660±21	666±31	721±42	656±49
Han		312	304±10	324±14	332±23	364±25	332±36
Papuan		252	256±9	273±13	267±17	331±34	314±55

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Expected amplitudes were computed according to formulas (2.10) and (2.11). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table 2.4. Amplitudes of weighted LD curves (multiplied by 10^6) for simulated 90% YRI / 10% CEU mixtures.

Ref 1	Ref 2	Expected	10 gen	20 gen	50 gen	100 gen	200 gen
Yoruba	French	563	587±27	579±26	550±25	600±43	562±96
Yoruba	Han	333	353±20	336±15	339±17	381±49	456±128
Yoruba	Papuan	289	307±19	303±16	309±18	343±54	426±248
San	French	488	522±25	512±22	488±25	519±28	625±89
San	Han	276	305±18	291±12	289±16	338±23	464±132
San	Papuan	236	266±18	262±13	254±12	306±38	486±186
Yoruba		6	6±1	6±1	7±1	7±3	44±89
San		1	16±15	8±3	10±7	-0±0	-1±1
French		454	473±19	471±18	450±19	481±19	566±55
Han		250	268±13	261±10	264±11	288±23	369±68
Papuan		212	231±14	233±13	243±11	276±35	366±125

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various references. Rows in which only one reference is listed indicate runs using the admixed population itself as one reference. Expected amplitudes were computed according to formulas (2.10) and (2.11). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table 2.5. Mixture fraction lower bounds on simulated 75% YRI / 25% CEU mixtures.

Ref	10	20	50	100	200
French	24.6±0.3	25.7±0.5	25.7±0.7	27.0±1.0	25.2±1.3
Russian	23.8±0.3	24.9±0.5	24.8±0.7	25.6±0.8	25.3±1.0
Sardinian	21.3±0.3	21.9±0.5	22.0±0.6	23.6±0.9	22.3±1.1
Kalash	14.7±0.2	15.5±0.4	15.5±0.5	16.4±0.6	15.6±0.9
Yoruba	73.6±0.7	74.8±0.4	74.0±0.6	76.2±1.3	73.8±3.4
Mandenka	50.5±0.6	51.2±1.0	50.4±1.4	54.9±2.0	60.8±5.6

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various single references. The first four rows are European surrogates and give lower bounds on the amount of CEU ancestry (25%); the last two are African surrogates and give lower bounds on the amount of YRI ancestry (75%). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table 2.6. Mixture fraction lower bounds on simulated 90% YRI / 10% CEU mixtures.

Ref	10	20	50	100	200
French	10.5±0.4	10.5±0.3	9.9±0.3	10.6±0.4	12.3±1.0
Russian	10.2±0.3	10.0±0.3	9.7±0.3	10.3±0.5	11.8±0.9
Sardinian	9.3±0.3	9.2±0.3	8.7±0.3	9.5±0.4	10.3±1.2
Kalash	7.2±0.3	7.0±0.3	6.8±0.2	7.4±0.4	8.9±0.8
Yoruba	89.1±1.0	89.1±1.1	90.1±1.5	89.4±3.7	98.5±2.0
Mandenka	18.2±2.3	17.3±2.5	19.5±4.8	63.1±25.5	30.7±220.4

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using various single references. The first four rows are European surrogates and give lower bounds on the amount of CEU ancestry (10%); the last two are African surrogates and give lower bounds on the amount of YRI ancestry (90%). Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

more accurate, with smaller standard errors. By contrast, we observed that while using a very small sample size (say 5) for the reference populations does create noticeable noise, using 20 samples already gives allele frequency estimates accurate enough that adding more reference samples has only minimal effects on the performance of *ALDER*. This is similar to the phenomenon that the precision of f-statistics does not improve appreciably with more than a moderate number of samples and is due to the inherent variability in genetic drift among different loci.

2.3.2 Robustness

A challenge of weighted LD analysis is that owing to various kinds of model violation, the parameters of the exponential fit of an observed curve $\hat{a}(d)$ may depend on the starting distance d_0 from which the curve is fit. We therefore explored the robustness of the fitting parameters to the choice of d_0 in a few scenarios (Figure 2.6). First, in a simulated 75% / 25% YRI–CEU admixture 50 generations ago, we find that the decay constant and amplitude are both highly robust to varying d_0 from 0.5 to 2.0 cM (Figure 2.6, top). This result is not surprising because our simulated example represents a true point admixture with minimal background LD in the admixed population.

In practice, we expect some dependence on d_0 due to background LD or longer-term admixture (either continuously over a stretch of time or in multiple waves). Both of these will tend to increase the weighted LD for smaller values of d relative to an exact exponential curve, so that estimates of the decay constant and amplitude will decrease as we increase the fitting start point d_0; the extent to which this effect occurs will depend on the extent of the model violation. We studied the d_0-dependence for two example admixed populations, HGDP Uygur and HapMap Maasai (MKK). For Uygur, the estimated decay constants and amplitudes are fairly robust to the start point of the fitting, varying roughly by ±10% (Figure 2.6, middle). In contrast, the estimates for Maasai vary dramatically, decreasing by more than a factor of 2 as d_0 is increased from 0.5 to 2.0 cM (Figure 2.6, bottom). This

Table 2.7. Dates of admixture estimated for simulated 75% YRI / 25% CEU mixtures.

Yoruba–French references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	12±2	18±2	55±3	103±7	258±24
10	10±1	19±2	50±2	105±7	236±24
20	10±1	20±1	52±2	104±5	223±16
50	9±0	20±1	52±1	96±2	186±10
100	10±0	20±0	52±1	101±2	210±9
San–Han references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	12±2	18±2	58±5	107±11	283±73
10	10±1	19±2	54±3	114±8	219±64
20	10±1	21±1	55±2	115±6	219±46
50	9±0	21±1	54±1	107±5	213±20
100	9±0	21±1	53±1	105±5	216±13

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using varying numbers of admixed samples. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

Table 2.8. Dates of admixture estimated for simulated 90% YRI / 10% CEU mixtures.

Yoruba–French references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	11±2	21±2	52±6	101±17	253±42
10	11±1	19±1	48±4	94±8	241±46
20	11±1	21±1	48±3	102±8	209±30
50	11±0	21±1	48±2	98±5	202±21
100	10±0	20±1	50±1	99±4	185±15
San–Han references					
Samples	10 gen	20 gen	50 gen	100 gen	200 gen
5	14±2	22±3	63±8	110±30	335±91
10	12±1	20±2	54±4	110±15	265±55
20	12±1	21±1	52±4	131±15	234±33
50	11±0	20±1	53±4	122±8	221±23
100	11±0	20±0	53±3	109±5	219±10

We simulated scenarios in which admixture occurred 10, 20, 50, 100, or 200 generations ago and show results from runs of *ALDER* using varying numbers of admixed samples. Note that standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation (not standard errors from averaging over multiple simulations).

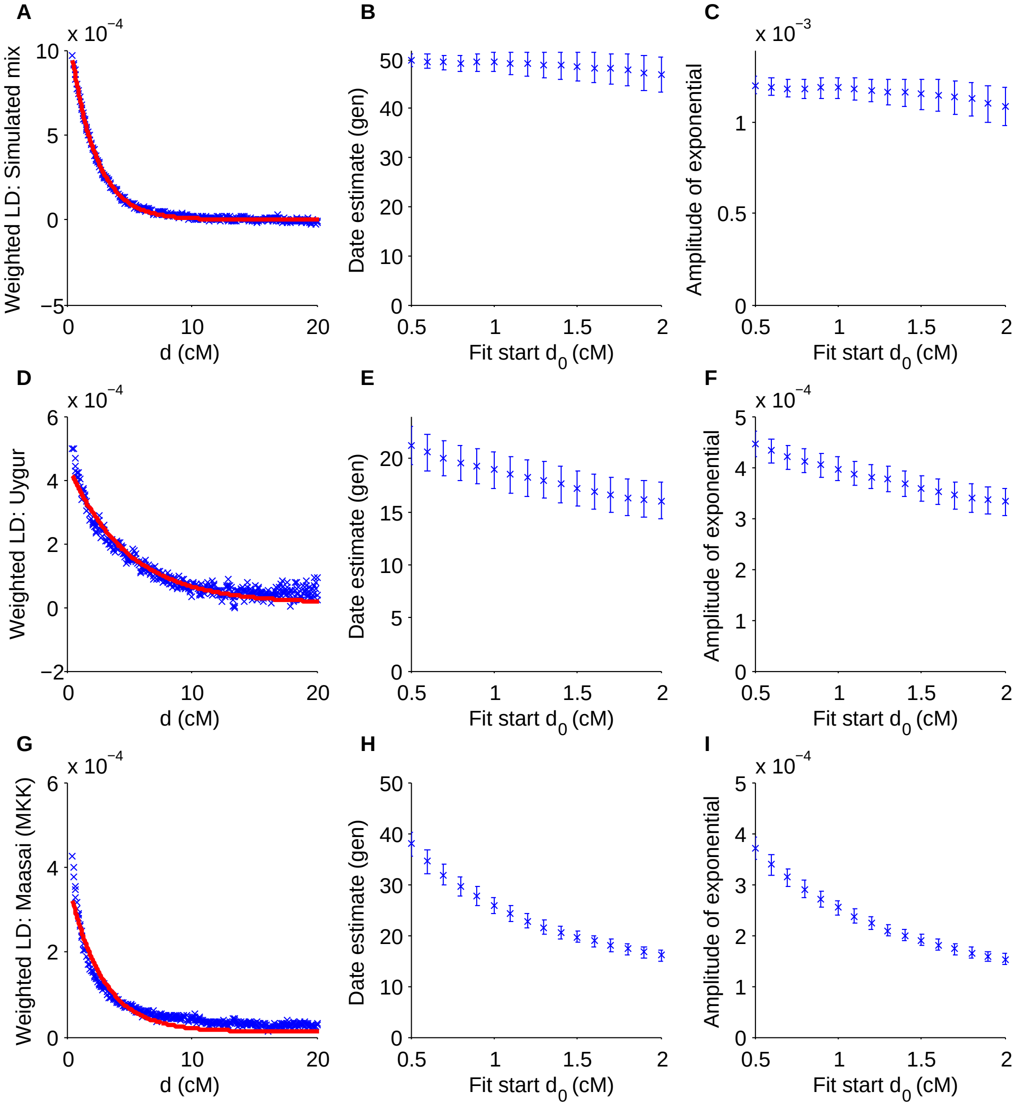


Figure 2.6. Dependence of date estimates and weighted LD amplitudes on fitting start point. Rows correspond to three test scenarios: Simulated 75% YRI / 25% CEU mixture 50 generations ago with Yoruba–French weights (top); Uygur with Han–French weights (middle); HapMap Maasai with Yoruba–French weights (bottom). The left panel of each row shows the weighted LD curve $\hat{a}(d)$ (blue) with best-fit exponential decay curve (red), fit starting from $d_0 = 0.5$ cM. Remaining panels show the date estimate (middle) and amplitude (right) as a function of d_0. (We note that our date estimates for Uygur are somewhat more recent than those in Patterson et al. (2012), most likely because of our direct estimate of the affine term in the weighted LD curve.)

Table 2.9. Effect of SNP ascertainment on date estimates.

Mixed pop	Ref 1	Ref 2	French asc	Han asc	San asc	Yoruba asc
Burusho	French	Han	47±12	51±13	56±10	41±10
Uygur	French	Han	15±2	14±2	13±2	16±2
Hazara	French	Han	22±2	22±3	23±2	22±3
Melanesian	Dai	Papuan	93±24	62±15	76±13	70±18
Bedouin	French	Yoruba	27±3	23±3	23±3	24±3
MbutiPygmy	San	Yoruba	33±12	33±6	41±14	30±8
BiakaPygmy	San	Yoruba	39±6	50±14	35±6	36±7

We compared dates of admixture estimated by *ALDER* on a variety of test triples from the HGDP using SNPs ascertained as heterozygous in full genome sequences of one French, Han, San, and Yoruba individual (Panels 1, 2, 4, and 5 of the Affymetrix Human Origins Array (Patterson et al., 2012)). Standard errors are from a jackknife over the 22 autosomes.

behavior is likely due to multiple-wave admixture in the genetic history of the Maasai; indeed, it is visually evident that the weighted LD curve for Maasai deviates from an exponential fit (Figure 2.6) and is in fact better-fit as a sum of exponentials. (See Figure 2.7 and Appendix B.3 for further simulations exploring continuous admixture.)

It is also important to consider the possibility of SNP ascertainment bias, as in any study based on allele frequencies. We believe that for weighted LD, ascertainment bias could have modest effects on the amplitude, which depends on F_2 distances (Patterson et al., 2012; Lipson et al., 2012), but will not affect the estimated date. Running *ALDER* on a suite of admixed populations in the HGDP under a variety of ascertainment schemes suggests that admixture date estimates are indeed quite stable to ascertainment (Table 2.9). Meanwhile, the amplitudes of the LD curves can scale substantially when computed under different SNP ascertainments, but their relative values are only different for extreme cases of African vs. non-African test populations under African vs. non-African ascertainment (Table 2.10; cf. Table 2 of Patterson et al. (2012)).

2.3.3 Admixture test results for HGDP populations

To compare the sensitivity of our LD-based test for admixture to the f-statistic-based 3-population test, we ran both *ALDER* and the 3-population test on all triples of populations in the HGDP. Interestingly, while the tests concur on the majority of the populations they identify as admixed, each also identifies several populations as admixed that the other does not (Table 2.11), showing that the tests have differing sensitivity to different admixture scenarios.

Admixture identified only by *ALDER*

The 3-population test loses sensitivity primarily as a result of drift since splitting from the references' lineages. More precisely, using the notation of Figure 2.1, the 3-population test statistic $f_3(C; A', B')$ estimates the sum of two directly competing terms: $-\alpha\beta F_2(A'', B'')$,

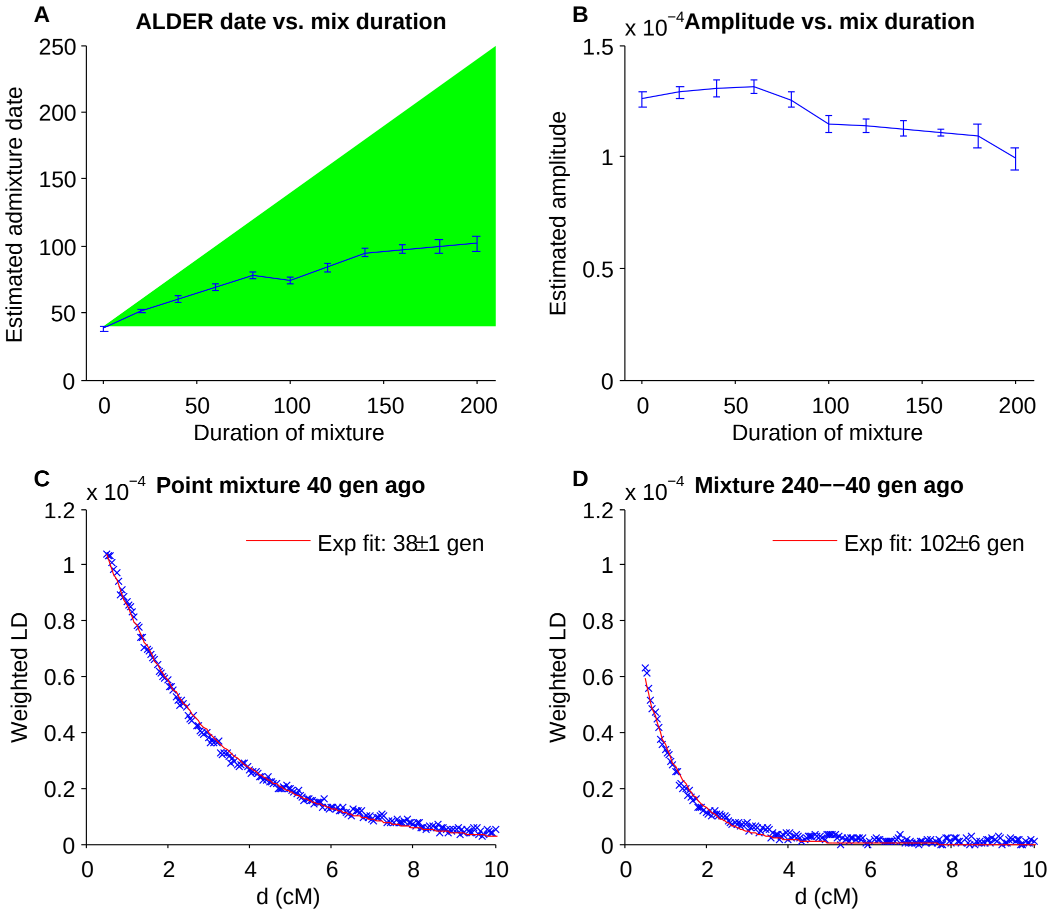


Figure 2.7. Weighted LD curve parameters from coalescent simulations of continuous admixture. In each simulation the mixed population receives 40% of its ancestry through continuous gene flow over a period of 0–200 generations ending 40 generations ago. Panels (A) and (B) show the admixture dates and weighted LD amplitudes computed by *ALDER* for each of 11 simulations (varying the duration of mixture from 0 to 200 in increments of 20). Panels (C) and (D) show the curves and exponential fits for mixture durations at the two extremes. Standard errors shown are *ALDER*'s jackknife estimates of its own error on a single simulation.

Table 2.10. Effect of SNP ascertainment on weighted LD curve amplitudes (multiplied by 10^6).

Mixed pop	Ref 1	Ref 2	French asc	Han asc	San asc	Yoruba asc
Burusho	French	Han	180±44	171±53	61±11	65±15
Uygur	French	Han	360±28	304±29	102±7	161±19
Hazara	French	Han	442±31	436±48	146±10	203±21
Melanesian	Dai	Papuan	868±277	559±150	207±51	312±91
Bedouin	French	Yoruba	227±32	196±25	104±11	146±13
MbutiPygmy	San	Yoruba	64±23	78±14	83±26	82±18
BiakaPygmy	San	Yoruba	104±19	133±46	90±15	103±22

We compared amplitudes of weighted LD curves fitted on a variety of test triples from the HGDP using SNPs ascertained as heterozygous in full genome sequences of one French, Han, San, and Yoruba individual (Panels 1, 2, 4, and 5 of the Affymetrix Human Origins Array (Patterson et al., 2012)). Standard errors are from a jackknife over the 22 autosomes.

the negative quantity arising from admixture that we wish to detect, and $\alpha^2 F_2(A'', A) + \beta^2 F_2(B'', B) + F_2(C, C')$, a positive quantity from the "off-tree" drift branches. If the latter term dominates, the 3-population test will fail to detect admixture regardless of the statistical power available. For example, Melanesians are only found to be admixed according to the *ALDER* test; the inability of the 3-population test to identify them as admixed is likely due to long off-tree drift from the Papuan branch prior to admixture. The situation is similar for the Pygmies, for whom we do not have two close references available.

Small mixture fractions also diminish the size of the admixture term $-\alpha\beta F_2(A, B)$ relative to the off-tree drift, and we believe this effect along with post-admixture drift may be the reason Sardinians are detected as admixed only by *ALDER*. In the case of the San, who have a small amount of Bantu admixture (Pickrell et al., 2012), the small mixture fraction may again play a role, along with the lack of a reference population closely related to the pre-admixture San, meaning that using existing populations incurs long off-tree drift.

Admixture identified only by the 3-population test

There are also multiple reasons why the 3-population test can identify admixture when *ALDER* does not. For the HGDP European populations in this category (Table 2.11), the 3-population test is picking up a signal of admixture identified by Patterson et al. (2012) and interpreted there as a large-scale admixture event in Europe involving Neolithic farmers closely related to present-day Sardinians and an ancient northern Eurasian population. This mixture likely began quite anciently (e.g. 7,000-9,000 years ago when agriculture arrived in Europe (Bramanti et al., 2009; Soares et al., 2010; Pinhasi et al., 2012)), and because admixture LD breaks down as e^{-nd}, where n is the age of admixture, there is nearly no LD left for *ALDER* to harness beyond the correlation threshold d_0. An additional factor that may inhibit LD-based testing is that in order to prevent false-positive identifications of admixture, *ALDER* typically eliminates reference populations that share LD (and in particular, admixture history) with the test population, whereas the 3-population test can

Table 2.11. Results of *ALDER* and 3-population tests for admixture on HGDP populations.

Both	#LD	# f_3	Only LD	#	Only f_3	#	Neither
Adygei	205	139	BiakaPygmy	81	French	99	Basque
Balochi	123	204	Colombian	5	Han	13	Dai
BantuKenya	30	182	Druze	128	Italian	46	Hezhen
BantuSouthAfrica	27	11	Japanese	1	Orcadian	1	Karitiana
Bedouin	300	63	Kalash	20	Tujia	8	Lahu
Brahui	363	16	MbutiPygmy	77	Tuscan	59	Mandenka
Burusho	450	377	Melanesian	96			Miao
Cambodian	266	158	Pima	489			Naxi
Daur	29	8	San	155			Papuan
Han-NChina	1	77	Sardinian	45			She
Hazara	699	593	Yakut	435			Surui
Makrani	173	163					Yi
Maya	784	124					Yoruba
Mongola	76	385					
Mozabite	313	107					
Oroqen	68	5					
Palestinian	308	64					
Pathan	113	348					
Russian	158	153					
Sindhi	264	366					
Tu	22	315					
Uygur	428	616					
Xibo	101	335					

We ran both *ALDER* and the 3-population test for admixture on each of the 53 HGDP populations using all pairs of other populations as references. We group the populations according to whether or not each test methodology produced at least one test identifying them as admixed; for each population, we list the number of reference pairs with which with each method (abbreviated "LD" and "f_3") detected admixture. We used a significance threshold of $p < 0.05$ after multiple-hypothesis correction.

use such references.

To summarize, the *ALDER* and 3-population tests both analyze a test population for admixture using two references, but they detect signal based on different “genetic clocks.” The 3-population test uses signal from genetic drift, which can detect quite old admixture but must overcome a counteracting contribution from post-admixture and off-tree drift. The LD-based test uses recombination, which is relatively unaffected by small population size-induced long drift and has no directly competing effect, but has limited power to detect chronologically old admixtures because of the rapid decay of the LD curve. Additionally, as discussed above in the context of simulation results, the LD-based test may be better suited for large data sets, since its power is enhanced more by the availability of many samples. The tests are thus complementary and both valuable. (See Figure 2.8 and Appendix B.3 for further exploration.)

2.3.4 Case studies

We now present detailed results for several human populations, all of which *ALDER* identifies as admixed but are not found by the 3-population test (Table 2.11). We infer dates of admixture and in some cases gain additional historical insights.

Pygmies

Both Central African Pygmy populations in the HGDP, the Mbuti and Biaka, show evidence of admixture (Table 2.11), about 28 ± 4 generations (800 years) ago for Mbuti and 38 ± 4 generations (1100 years) ago for Biaka, estimated using San and Yoruba as reference populations (Figure 2.9A,C). The intra-population heterogeneity is low, as demonstrated by the negligible affine terms. In each case, we also generated weighted LD curves with the Pygmy population itself as one reference and a variety of second references. We found that using French, Han, or Yoruba as the second reference gave very similar amplitudes, but the amplitude was significantly smaller with the other Pygmy population or San as the second reference (Figure 2.9B,D). Using the amplitudes with Yoruba, we estimated mixture fractions of at least $15.9 \pm 0.9\%$ and $28.8 \pm 1.4\%$ Yoruba-related ancestry (lower bounds) for Mbuti and Biaka, respectively.

The phylogenetic interpretation of the relative amplitudes is complicated by the fact that the Pygmy populations, used as references, are themselves admixed, but a plausible coherent explanation is as follows (see Figure 2.9E). We surmise that a proportion β (bounds given above) of Bantu-related gene flow reached the native Pygmy populations on the order of 1000 years ago. The common ancestors of Yoruba or non-Africans with the Bantu population are genetically not very different from Bantu, due to high historical population sizes (branching at positions X_1 and X_2 in Figure 2.9E). Thus, the weighted LD amplitudes using Yoruba or non-Africans as second references are nearly $2\alpha^3\beta F_2(A,B)^2$, where B denotes the admixing Bantu population. Meanwhile, San and Western (resp. Eastern) Pygmies split from the Bantu-Mbuti (resp. Biaka) branch toward the middle or the opposite side from Bantu (X_3 and X_4), giving a smaller amplitude (Figure 2.3).

Our results are in agreement with previous studies that have found evidence of gene flow from agriculturalists to Pygmies (Quintana-Murci et al., 2008; Verdu et al., 2009; Patin

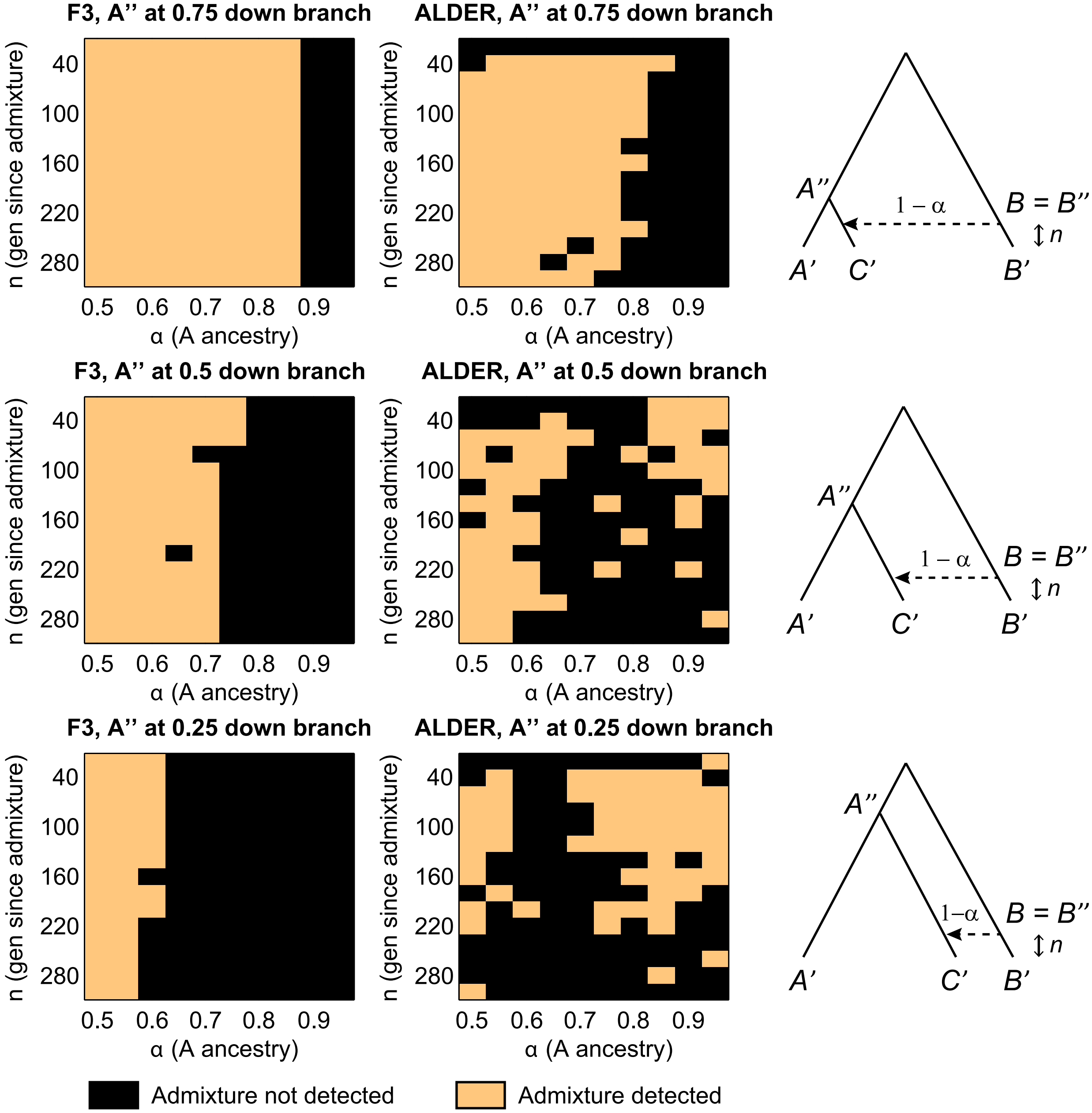


Figure 2.8. Coalescent simulations comparing the sensitivities of the 3-population moment-based test for admixture (f_3) and the LD-based test implemented in *ALDER*. We varied three parameters: the age of the branch point A'', the date n of gene flow, and the fraction α of A ancestry.

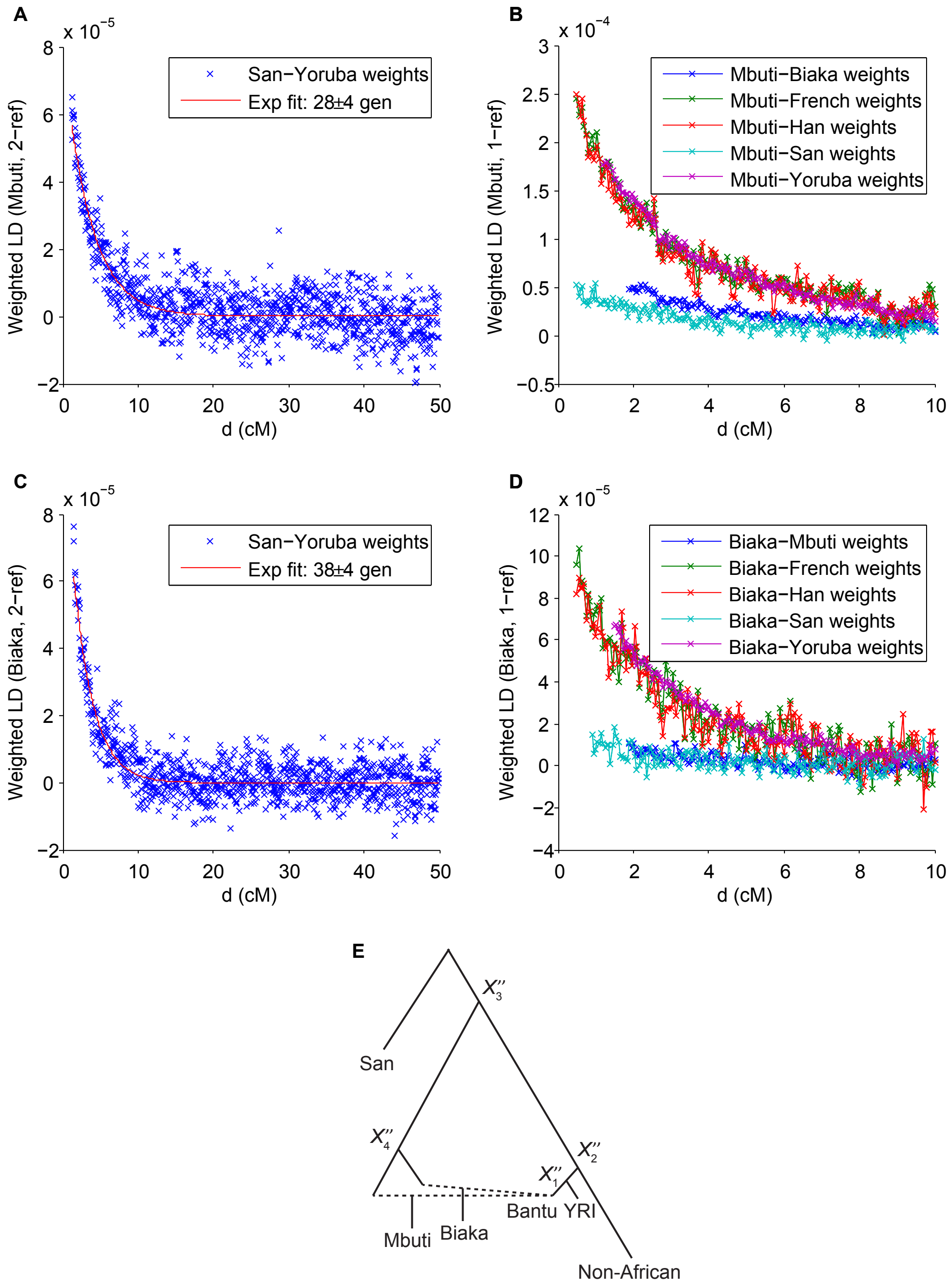


Figure 2.9. Weighted LD curves for Mbuti using San and Yoruba as reference populations (A) and using Mbuti itself as one reference and several different second references (B), and analogous curves for Biaka (C, D). Genetic distances are discretized into bins at 0.05 cM resolution. Data for each curve are plotted and fit starting from the corresponding *ALDER*-computed LD correlation thresholds. Different amplitudes of one-reference curves (B, D) imply different phylogenetic positions of the references relative to the true mixing populations (i.e., different split points X_i''), suggesting a sketch of a putative admixture graph (E). Relative branch lengths are qualitative, and the true root is not necessarily as depicted.

Table 2.12. Amplitudes and dates from weighted LD curves for Sardinian using various reference pairs.

Ref 1	Ref 2	Weighted LD amplitude	Date estimate
CEU	YRI	0.00003192 ± 0.00000903	48 ± 10
CHB	YRI	0.00001738 ± 0.00000679	34 ± 8
CEU	CHB	0.00000873 ± 0.00000454	52 ± 21

Data are shown from *ALDER* fits to weighted LD curves computed using Sardinian as the test population and pairs of HapMap CEU, YRI, and CHB as the references. Date estimates are in generations. We omitted chromosome 8 from the analysis because of anomalous long-range LD. Curves $\hat{a}(d)$ were fit for $d > 1.2$ cM, the extent of LD correlation between Sardinian and CEU computed by *ALDER*.

et al., 2009; Jarvis et al., 2012). Quintana-Murci et al. (2008) suggested based on mtDNA evidence in Mbuti that gene flow ceased several thousand years ago, but more recently, Jarvis et al. (2012) found evidence of admixture in Western Pygmies, with a local-ancestry-inferred block length distribution of 3.1 ± 4.6 Mb (mean and standard deviation), consistent with our estimated dates.

Sardinians

We detect a very small proportion of Sub-Saharan African ancestry in Sardinians, which our *ALDER* tests identified as admixed (Table 2.11; Figure 2.10A). To investigate further, we computed weighted LD curves with Sardinian as the test population and all pairs of the HapMap CEU, YRI and CHB populations as references (Table 2.12). We observed an abnormally large amount of shared long-range LD in chromosome 8, likely do to an extended inversion segregating in Europeans (Price et al., 2008), so we omitted it from these analyses. The CEU–YRI curve has the largest amplitude, suggesting both that the LD present is due to admixture and that the small non-European ancestry component, for which we estimated a lower bound of $0.6 \pm 0.2\%$, is from Africa. (For this computation we used single-reference weighted LD with YRI as the reference, fitting the curve after 1.2 cM to reduce confounding effects from correlated LD that *ALDER* detected between Sardinian and CEU. Changing the starting point of the fit does not qualitatively affect the results.) The existence of a weighted LD decay curve with CHB and YRI as references provides further evidence that the LD is not simply due to a population bottleneck or other non-admixture sources, as does the fact that our estimated dates from all three reference pairs are roughly consistent at about 40 generations (1200 years) ago. Our findings thus confirm the signal of African ancestry in Sardinians reported in Moorjani et al. (2011). The date, small mixture proportion, and geography are consistent with a small influx of migrants from North Africa, who themselves traced only a fraction of their ancestry ultimately to Sub-Saharan Africa, consistent with the findings of Dupanloup et al. (2004).

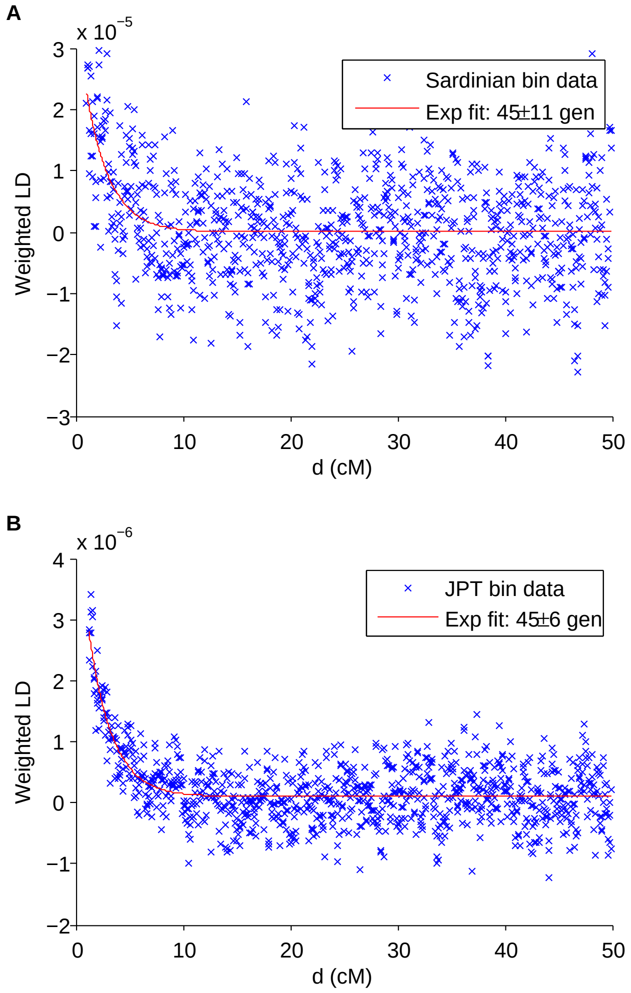


Figure 2.10. Weighted LD curves for HGDP Sardinian using Italian–Yoruba weights (A) and HapMap Japanese (JPT) using JPT itself as one reference and HapMap Han Chinese (CHB) as the second reference (B). The exponential fits are performed starting at 1 cM and 1.2 cM, respectively, as selected by *ALDER* based on detected correlated LD.

Japanese

Genetic studies have suggested that present-day Japanese are descended from admixture between two waves of settlers, responsible for the Jomon and Yayoi cultures (Hammer and Horai, 1995; Hammer et al., 2006; Rasteiro and Chikhi, 2009). We also observed evidence of admixture in Japanese, and while our ability to learn about the history was limited by the absence of a close surrogate for the original Paleolithic mixing population, we were able to take advantage of the one-reference inference capabilities of *ALDER*. More precisely, among our tests using all pairs of HGDP populations as references (Table 2.11), one reference pair, Basque and Yakut, produced a passing test for Japanese. However, as we have noted, the reference populations need not be closely related to the true mixing populations, and we believe that in this case this seemingly odd reference pair arises as the only passing test because the data set lacks a close surrogate for Jomon.

In the absence of a reference on the Jomon side, we computed single-reference weighted LD using HapMap JPT as the test population and JPT–CHB weights, which confer the advantage of larger sample sizes (Figure 2.10B). The weighted LD curve displays a clear decay, yielding an estimate of 45 ± 6 generations, or about 1,300 years, as the age of admixture. To our knowledge, this is the first time genome-wide data have been used to date admixture in Japanese. As with previous estimates based on coalescence of Y-chromosome haplotypes (Hammer et al., 2006), our date is consistent with the archaeologically attested arrival of the Yayoi in Japan roughly 2300 years ago (we suspect that our estimate is from later than the initial arrival because admixture may not have happened immediately or may have taken place over an extended period of time). Based on the amplitude of the curve, we also obtain a (likely very conservative) genome-wide lower bound of $41 \pm 3\%$ "Yayoi" ancestry using formula (2.12) (under the reasonable assumption that Han Chinese are fairly similar to the Yayoi population). It is important to note that the observation of a single-reference weighted LD curve is not sufficient evidence to prove that a population is admixed, but the existence of a pair of references with which the *ALDER* test identified Japanese as admixed, combined with previous work and the lack of any signal of reduced population size, makes us confident that our inferences are based on true historical admixture.

Onge

Lastly, we provide a cautionary example of weighted LD decay curves arising from demography and not admixture. We observed distinct weighted LD curves when analyzing the Onge, an indigenous population of the Andaman Islands. However, this curve is only present when using Onge themselves as one reference; moreover, the amplitude is independent of whether CEU, CHB, YRI, GIH (HapMap Gujarati), or Great Andamanese is used as the second reference (Figure 2.11), as expected if the weighted LD is due to correlation between LD and allele frequencies in the test population alone (and independent of the reference allele frequencies). Correspondingly, *ALDER*'s LD-based test does not identify Onge as admixed using any pair of these references. Thus, while we cannot definitively rule out admixture, the evidence points toward internal demography (low population size) as the cause of the elevated LD, consistent with the current census of fewer than 100 Onge individuals.

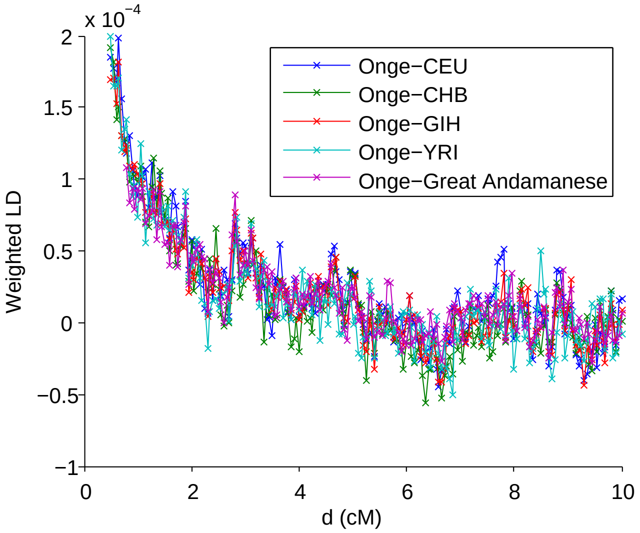


Figure 2.11. Weighted LD curves for Onge using Onge itself as one reference and several different second references.

2.4 Discussion

2.4.1 Strengths of weighted LD for admixture inference

The statistics underlying weighted LD are quite simple, making the formula for the expectation of $\hat{a}(d)$, as well as the noise and other errors from our inference procedure, relatively easy to understand. By contrast, local ancestry-based admixture dating methods (e.g., Pool and Nielsen (2009) and Gravel (2012)) are sensitive to imperfect ancestry inference, and it is difficult to trace the error propagation to understand the ultimate effect on inferred admixture parameters. Similarly, the wavelet method of Pugach et al. (2011) uses reference populations to perform (fuzzy) ancestry assignment in windows, for which error analysis is challenging.

Another strength of our weighted LD methodology is that it has relatively low requirements on the quality and quantity of reference populations. Our theory tells us exactly how the statistic behaves for any reference populations, no matter how diverged they are from the true ancestral mixing populations. In contrast, the accuracy of results from clustering and local ancestry methods is dependent on the quality of the reference populations used in ways that are difficult to characterize. On the quantity side, previous approaches to admixture inference require a surrogate for each ancestral population, whereas as long as one is confident that the signal is truly from admixture, weighted LD can be used with only one available reference to infer times of admixture (as in our analysis of the Japanese) and bound mixing fractions (as in our Pygmy case study and Pickrell et al. (2012)), problems that were previously intractable.

Weighted LD also advances our ability to test for admixture. As discussed above, *ALDER* offers complementary sensitivity to the 3-population test and allows the identification of additional populations as admixed. Another formal test for admixture is the 4-population test (Reich et al., 2009; Patterson et al., 2012), which is quite sensitive but also has tradeoffs; for example, it requires three distinctly branching references, whereas *ALDER* and the 3-population test only need two. Additionally, the phylogeny of the populations involved must be well understood in order to interpret a signal of admixture from the 4-population test properly (i.e., to determine which population is admixed). Using weighted LD, on the other hand, largely eliminates the problem of determining the destination or direction of gene flow, since the LD signal of admixture is intrinsic to a specified test population.

2.4.2 One-reference versus two-reference curves

In practice, it is often useful to compute weighted LD curves using both the one-reference and two-reference techniques, as both can be used for inferences in different situations. Generally, we consider two-reference curves to be more reliable for parameter estimation, since using the test population as one reference is more prone to introduce unwanted signals, such as recent admixture from a different source, non-admixture LD from reduced population size, or population structure among samples. In particular, populations with more complicated histories and additional sources of LD beyond the specifications of our model will often have different estimates of admixture dates with one- and two-reference curves. There is a small chance that date disagreement can reflect a false-positive admixture signal, but this is very unlikely if both one- and two-reference curves exist beyond the correlated LD threshold (see Appendix B.2). Two-reference curves also allow for direct estimation of mixture fractions, although, as discussed above, we prefer instead to use the method of single-reference bounding.

There are a number of practical considerations that make the one-reference capabilities of *ALDER* desirable. Foremost is the possibility that one may not have a good surrogate available for one of the ancestral mixing populations, as in our Japanese example. Also, while our method of learning about phylogenetic relationships is best suited to two-reference curves because of the simpler form of the amplitude in terms of branch lengths, it is often useful to begin by computing a suite of single-reference curves, both because the data generated will scale linearly with the number of references available and because observing a range of different amplitudes gives an immediate signal of the presence of admixture in the test population.

Overall, then, a sample sequence for applying *ALDER* to a new data set might be as follows: (1) test all populations for admixture using all pairs of references from among the other populations; (2) explore admixed populations of interest by comparing single-reference weighted LD curves; (3) learn more detail by analyzing selected two-reference curves alongside the one-reference ones; (4) estimate parameters using one- or two-reference curves as applicable. Of course, step (1) itself involves the complementary usefulness of both one- and two-reference weighted LD, since our test for admixture requires the presence of exponential decay signals in both types of curves.

2.4.3 Effect of multiple-wave or continuous admixture

As discussed in our section on robustness of results, in the course of our data analysis, we observed that the weighted LD date estimate almost always becomes more recent when the exponential decay curve is fit for a higher starting distance d_0. Most likely, this is because admixtures in human populations have taken place over multiple generations, such that our estimated times represent intermediate dates during the process. To whatever extent an admixture event is more complicated than posited in our point-admixture model, removing low-d bins will lead the fitting to capture proportionately more of the more recent admixture. By default, *ALDER* sets d_0 to be the smallest distance such that non-admixture LD signals can be confidently discounted for $d > d_0$ (see Methods (Testing for admixture) and Appendix B.2), but it should be noted that the selected d_0 will vary for different sets of populations, and in each case the true admixture signal at $d < d_0$ will also be excluded. Theoretically, this pattern could allow us to learn more about the true admixture history of a population, since the value of $a(d)$ at each d represents a particular function of the amount of admixture that took place at each generation in the past. However, in our experience, fitting becomes difficult for any model involving more than two or three parameters. Thus, we made the decision to restrict ourselves to assuming a single point admixture, fit for a principled threshold $d > d_0$, accepting that the inferred date n represents some form of average value over the true history.

2.4.4 Other possible complications

In our derivations, we have assumed implicitly that the mixing populations and the reference populations are related through a simple tree. However, it may be that their history is more complicated, for example involving additional admixtures. In this case, our formulas for the amplitude of the ALD curve will be inaccurate if, for example, A and A' have different admixture histories. However, if our assumptions are violated only by events occurring before the divergences between the mixing populations and the corresponding references, then the amplitude will be unaffected. Moreover, no matter what the population history, as long as A and B are free of measurable LD (so that our assumption of independence of alleles conditional on a single ancestry is valid), there will be no effect on the estimated date of admixture.

2.4.5 Conclusions and future directions

In this study, we have shown how linkage disequilibrium (LD) generated by population admixture can be a powerful tool for learning about history, extending previous work that showed how it can be used for estimating dates of mixture (Moorjani et al., 2011; Patterson et al., 2012). We have developed a new suite of tools, implemented in the *ALDER* software package, that substantially increases the speed of admixture LD analysis, improves the robustness of admixture date inference, and exploits the amplitude of LD as a novel source of information about history. In particular, (a) we show how admixture LD can be leveraged into a formal test for mixture that can sometimes find evidence of admixture not detectable by other methods, (b) we show how to estimate mixture proportions, and (c) we show that we can even use this information to infer phylogenetic relationships. A limitation of *ALDER* at present, however, is that it is designed for a model of pulse admixture between two ancestral populations. Important directions for future work will be to generalize these ideas to make inferences about the time course of admixture in the case that it took place over a longer period of time (Pool and Nielsen, 2009; Gravel, 2012) and to study multi-way admixture. In addition, it would be valuable to be able to use the information from admixture LD to constrain models of history for multiple populations simultaneously, either by extending *ALDER* itself or by using LD-based test results in conjunction with methods for fitting phylogenies incorporating admixture (Patterson et al., 2012; Pickrell and Pritchard, 2012; Lipson et al., 2012).

2.5 Software

Executable and C++ source files for our *ALDER* software package are available online at the Berger and Reich Lab websites: `http://groups.csail.mit.edu/cb/alder/`, `http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html`.

Chapter 3

Admixture Inference Using Moment Statistics

The recent explosion in available genetic data has led to significant advances in understanding the demographic histories of and relationships among human populations. It is still a challenge, however, to infer reliable parameter values for complicated models involving many populations. Here we present *MixMapper*, an efficient, interactive method for constructing phylogenetic trees including admixture events using single nucleotide polymorphism (SNP) genotype data. *MixMapper* implements a novel two-phase approach to admixture inference using moment statistics, first building an unadmixed scaffold tree and then adding admixed populations by solving systems of equations that express allele frequency divergences in terms of mixture parameters. Importantly, all features of the model, including topology, sources of gene flow, branch lengths, and mixture proportions, are optimized automatically from the data and include estimates of statistical uncertainty. *MixMapper* also uses a new method to express branch lengths in easily interpretable drift units. We apply *MixMapper* to recently published data for HGDP individuals genotyped on a SNP array designed especially for use in population genetics studies, obtaining confident results for 30 populations, 20 of them admixed. Notably, we confirm a signal of ancient admixture in European populations—including previously undetected admixture in Sardinians and Basques—involving a proportion of 20–40% ancient northern Eurasian ancestry.[1]

3.1 Introduction

The most basic way to represent the evolutionary history of a set of species or populations is through a phylogenetic tree, a model that in its strict sense assumes that there is no gene flow between populations after they have diverged (Cavalli-Sforza and Edwards, 1967). In many settings, however, groups that have split from one another can still exchange genetic material. This is certainly the case for human population history, during the course of

[1] The material in this chapter was previously posted to the arXiv in December 2012 as "Efficient moment-based inference of admixture parameters and sources of gene flow" by Mark Lipson, Po-Ru Loh, Alex Levin, David Reich, Nick Patterson, and Bonnie Berger (Lipson et al., 2012) and is in revision at *Molecular Biology and Evolution* at the time of this writing.

which populations have often diverged only incompletely or diverged and subsequently mixed again (Reich et al., 2009; Wall et al., 2009; Laval et al., 2010; Green et al., 2010; Reich et al., 2010; Gravel et al., 2011; Patterson et al., 2012). To capture these more complicated relationships, previous studies have considered models allowing for continuous migration among populations (Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011) or have extended simple phylogenetic trees into *admixture trees*, in which populations on separate branches are allowed to re-merge and form an admixed offspring population (Chikhi et al., 2001; Wang, 2003; Reich et al., 2009; Sousa et al., 2009; Patterson et al., 2012). Both of these frameworks, of course, still represent substantial simplifications of true population histories, but they can help capture a range of new and interesting phenomena.

Several approaches have previously been used to build phylogenetic trees incorporating admixture events from genetic data. First, likelihood methods (Chikhi et al., 2001; Wang, 2003; Sousa et al., 2009) use a full probabilistic evolutionary model, which allows a high level of precision with the disadvantage of greatly increased computational cost. Consequently, likelihood methods can in practice only accommodate a small number of populations (Wall et al., 2009; Laval et al., 2010; Gravel et al., 2011; Sirén et al., 2011). Moreover, the tree topology must generally be specified in advance, meaning that only parameter values can be inferred automatically and not the arrangement of populations in the tree. By contrast, the moment-based methods of Reich et al. (2009) and Patterson et al. (2012) use only means and variances of allele frequency divergences. Moments are simpler conceptually and especially computationally, and they allow for more flexibility in model conditions. Their disadvantages can include reduced statistical power and difficulties in designing precise estimators with desirable statistical properties (e.g., unbiasedness) (Wang, 2003). Finally, a number of studies have considered “phylogenetic networks,” which generalize trees to include cycles and multiple edges between pairs of nodes and can be used to model population histories involving hybridization (Huson and Bryant, 2006; Yu et al., 2012). However, these methods also tend to be computationally expensive.

In this work, we introduce *MixMapper*, a new computational tool that fits admixture trees by solving systems of moment equations involving the pairwise distance statistic f_2 (Reich et al., 2009; Patterson et al., 2012), which is the average squared allele frequency difference between two populations. The theoretical expectation of f_2 can be calculated in terms of branch lengths and mixture fractions of an admixture tree and then compared to empirical data. *MixMapper* can be thought of as a generalization of the *qpgraph* package (Patterson et al., 2012), which takes as input genotype data, along with a proposed arrangement of admixed and unadmixed populations, and returns branch lengths and mixture fractions that produce the best fit to allele frequency moment statistics measured on the data. *MixMapper*, by contrast, performs the fitting in two stages, first constructing an unadmixed scaffold tree via neighbor-joining and then automatically optimizing the placement of admixed populations onto this initial tree. Thus, no topological relationships among populations need to be specified in advance.

Our method is similar in spirit to the independently developed *TreeMix* package (Pickrell and Pritchard, 2012). Like *MixMapper*, *TreeMix* builds admixture trees from second moments of allele frequency divergences, although it does so via a composite likelihood maximization approach made tractable with a multivariate normal approximation. Procedurally, *TreeMix* initially fits a full set of populations as an unadmixed tree, and gene flow edges are added sequentially to account for the greatest errors in the fit (Pickrell and Pritchard, 2012). This format makes *TreeMix* well-suited to handling very large trees: the entire fitting process is automated and can include arbitrarily many admixture events simultaneously. In contrast, *MixMapper* begins with a carefully screened unadmixed scaffold tree to which admixed populations are added with best-fitting parameter values, an interactive design that enables precise modeling of particular populations of interest.

We use *MixMapper* to model the ancestral relationships among 52 populations from the CEPH-Human Genome Diversity Cell Line Panel (HGDP) (Rosenberg et al., 2002; Li et al., 2008b) using recently published data from a new, specially ascertained SNP array designed for population genetics applications (Keinan et al., 2007; Patterson et al., 2012). Previous studies of these populations have built simple phylogenetic trees (Li et al., 2008b; Sirén et al., 2011), identified a substantial number of admixed populations with likely ancestors (Patterson et al., 2012), and constructed a large-scale admixture tree (Pickrell and Pritchard, 2012). Here, we add an additional level of quantitative detail, obtaining best-fit admixture parameters with bootstrap error estimates for 30 HGDP populations, of which 20 are admixed. The results include, most notably, a significant admixture event (Patterson et al., 2012) in the history of all sampled European populations, among them Sardinians and Basques.

3.2 New Approaches

The central problem we consider is: given an array of SNP data sampled from a set of individuals grouped by population, what can we infer about the admixture histories of these populations using simple statistics that are functions of their allele frequencies? Methodologically, the *MixMapper* workflow (Figure 3.1) proceeds as follows. We begin by computing f_2 distances between all pairs of study populations, from which we construct an unadmixed phylogenetic subtree to serve as a scaffold for subsequent mixture fitting. The choice of populations for the scaffold is done via initial filtering of populations that are clearly admixed according to the 3-population test (Reich et al., 2009; Patterson et al., 2012), followed by selection of a subtree that is approximately additive along its branches, as is expected in the absence of admixture (see Material and Methods and Appendix C.1 for full details).

Next, we expand the model to incorporate admixtures by attempting to fit each population not in the scaffold as a mixture between some pair of branches of the scaffold. Putative admixtures imply algebraic relations among f_2 statistics, which we test for consistency with the data, allowing us to identify likely sources of gene flow and estimate mixture parameters (Figure 3.2; Appendix C.1). After determining likely two-way admixture events, we further attempt to fit remaining populations as three-way mixtures involving the inferred two-way mixed populations, by similar means. Finally, we use a new formula to convert the f_2 tree distances into absolute drift units (Appendix C.2). Importantly, we apply a bootstrap resampling scheme (Efron, 1979; Efron and Tibshirani, 1986) to obtain ensembles of predictions, rather than single values, for all model variables. This procedure allows us to determine confidence intervals for parameter estimates and guard against overfitting. For a data set on the scale of the HGDP, after initial setup time on the order of an hour, *MixMapper* determines the best-fit admixture model for a chosen population in a few seconds, enabling real-time interactive investigation.

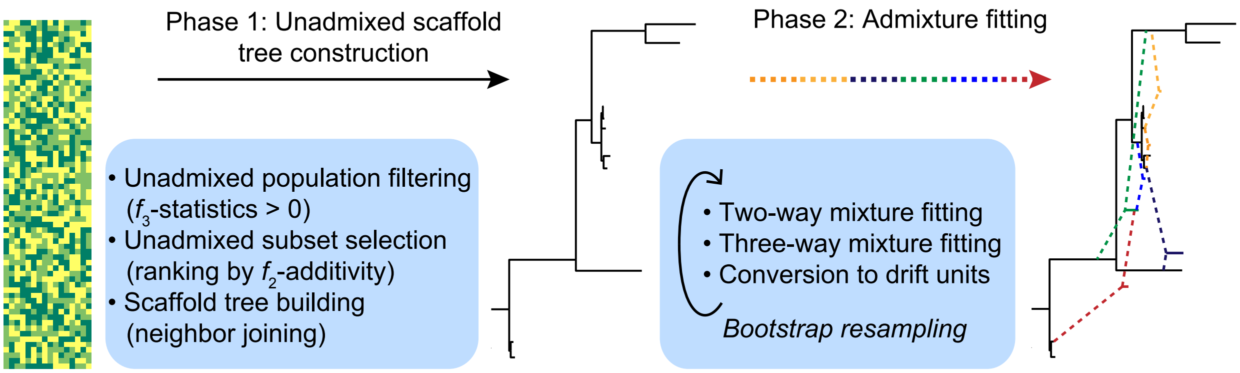

Phase 1: Unadmixed scaffold tree construction
• Unadmixed population filtering (f_3-statistics > 0)
• Unadmixed subset selection (ranking by f_2-additivity)
• Scaffold tree building (neighbor joining)
Phase 2: Admixture fitting
• Two-way mixture fitting
• Three-way mixture fitting
• Conversion to drift units
Bootstrap resampling

Figure 3.1. *MixMapper* workflow. *MixMapper* takes as input an array of SNP calls annotated with the population to which each individual belongs. The method then proceeds in two phases, first building a tree of (approximately) unadmixed populations and then attempting to fit the remaining populations as admixtures. In the first phase, *MixMapper* produces a ranking of possible unadmixed trees in order of deviation from f_2-additivity; based on this list, the user selects a tree to use as a scaffold. In the second phase, *MixMapper* tries to fit remaining populations as two- or three-way mixtures between branches of the unadmixed tree. In each case *MixMapper* produces an ensemble of predictions via bootstrap resampling, enabling confidence estimation for inferred results.

3.3 Results

3.3.1 Simulations

To test the inference capabilities of *MixMapper* on populations with known histories, we ran it on two data sets generated with the coalescent simulator `ms` (Hudson, 2002) and designed to have similar parameters to our human data. In both cases, we simulated 500 regions of 500 kb each for 25 diploid individuals per population, with an effective population size of 5,000 or 10,000 per population, a mutation rate of 0.5×10^{-8} per base per generation (intentionally low so as not to create unreasonably many SNPs), and a recombination rate of 10^{-8} per base per generation. Full `ms` commands can be found in Material and Methods. We ascertained SNPs present at minor allele frequency 0.05 or greater in an outgroup population and then removed that population from the analysis.

For the first admixture tree, we simulated six non-outgroup populations, with one of them, pop6, admixed (Figure 3.3A). Applying *MixMapper*, no admixtures were detected with the 3-population test, but the most additive subset with at least five populations excluded pop6 (max deviation from additivity 2.0×10^{-4} versus second-best 7.7×10^{-4}; see Material and Methods), so we used this subset as the scaffold tree. We then fit pop6 as admixed, and *MixMapper* recovered the correct gene flow topology with 100% confidence and inferred the other parameters of the model quite accurately (Figure 3.3B; Table 3.1). For comparison, we also analyzed the same data with *TreeMix* and again obtained accurate results (Figure 3.3C).

For the second test, we simulated a complex admixture scenario involving 10 non-outgroup populations, with six unadmixed and four admixed (Figure 3.3D). In this example,

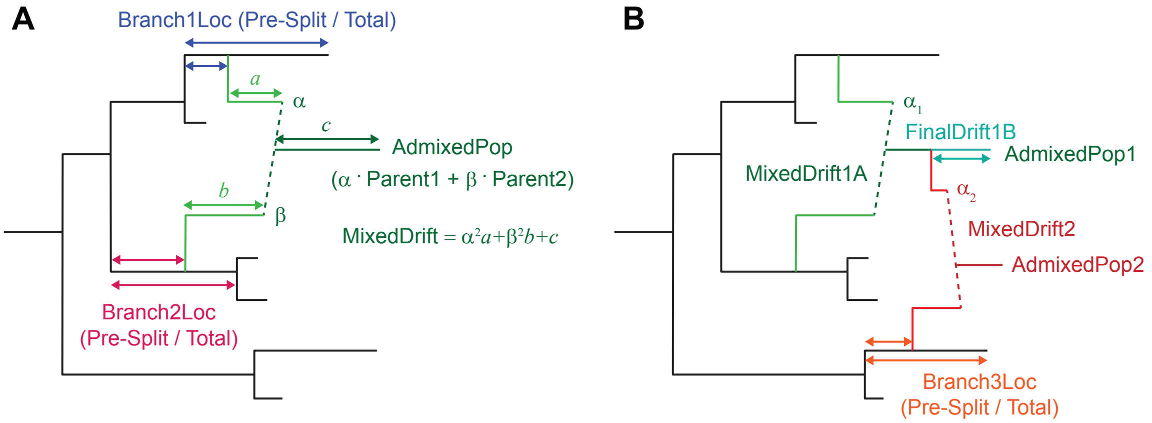


Figure 3.2. Schematic of mixture parameters fit by *MixMapper*. (A) A simple two-way admixture. *MixMapper* infers four parameters when fitting a given population as an admixture. It finds the optimal pair of branches between which to place the admixture and reports the following: Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches (given as pre-split length / total branch length); α is the proportion of ancestry from Branch1 ($\beta = 1 - \alpha$ is the proportion from Branch2); and MixedDrift is the linear combination of drift lengths $\alpha^2 a + \beta^2 b + c$. (B) A three-way mixture: here AdmixedPop2 is modeled as an admixture between AdmixedPop1 and Branch3. There are now four additional parameters; three are analogous to the above, namely, Branch3Loc, α_2, and MixedDrift2. The remaining degree of freedom is the position of the split along the AdmixedPop1 branch, which divides MixedDrift into MixedDrift1A and FinalDrift1B.

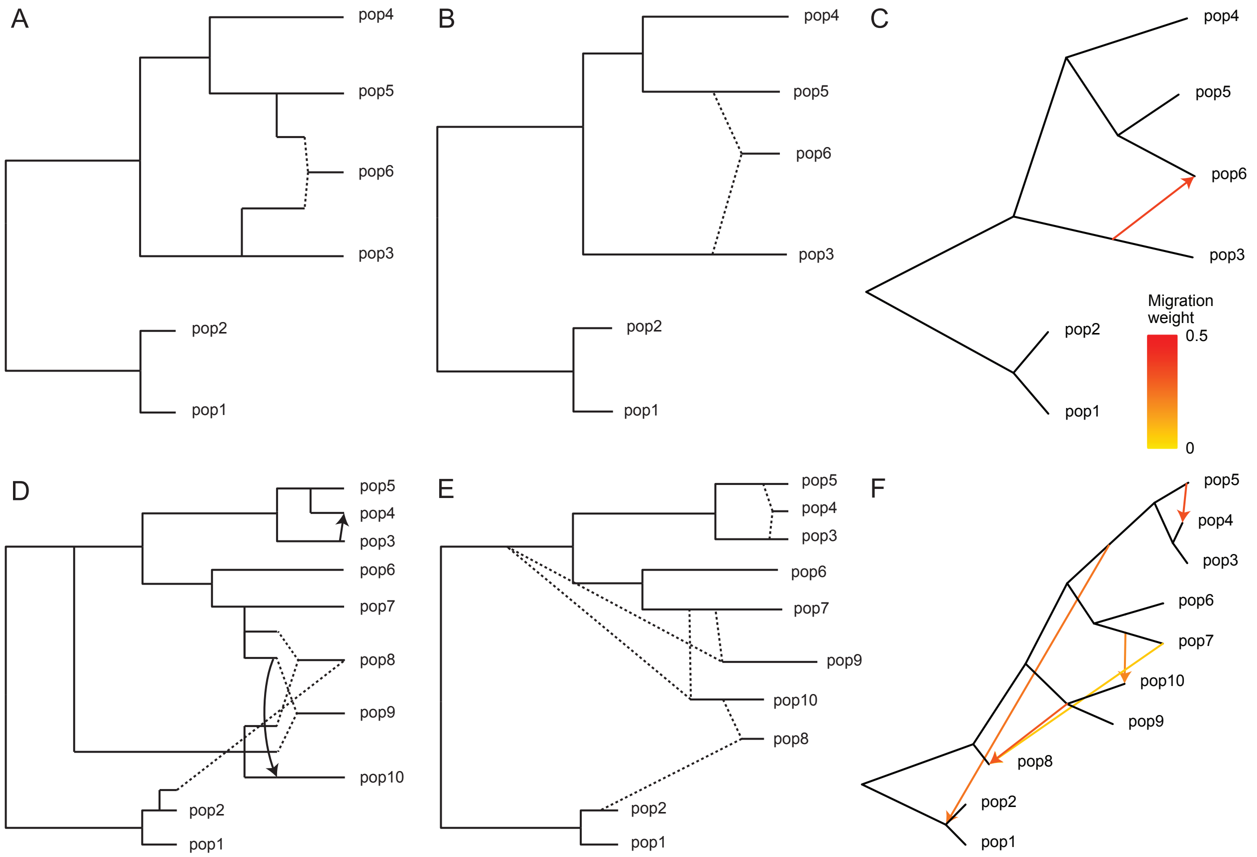


Figure 3.3. Results with simulated data. (A-C) First simulated admixture tree, with one admixed population. Shown are: (A) the true phylogeny, (B) *MixMapper* results, and (C) *TreeMix* results. (D-F) Second simulated admixture tree, with four admixed populations. Shown are: (D) the true phylogeny, (E) *MixMapper* results, and (F) *TreeMix* results. In (A) and (D), dotted lines indicate instantaneous admixtures, while arrows denote continuous (unidirectional) gene flow over 40 generations. Both *MixMapper* and *TreeMix* infer point admixtures, depicted with dotted lines in (B) and (E) and colored arrows in (C) and (F). In (B) and (E), the terminal drift edges shown for admixed populations represent half the total mixed drift. Full inferred parameters from *MixMapper* are given in Table 3.1.

Table 3.1. Mixture parameters for simulated data.

AdmixedPop	Branch1 + Branch2	# rep	α	Branch1Loc	Branch2Loc	MixedDrift
First tree						
pop6	pop3 + pop5	500	0.253-0.480	0.078-0.195 / 0.214	0.050-0.086 / 0.143	0.056-0.068
pop6 (true)	pop3 + pop5		0.4	0.107 / 0.213	0.077 / 0.145	0.066
Second tree						
pop4	pop3 + pop5	500	0.382-0.652	0.039-0.071 / 0.076	0.032-0.073 / 0.077	0.010-0.020
pop4 (true)	pop3 + pop5		0.4	0.071 / 0.077	0.038 / 0.077	0.016
pop9	Anc3–7 + pop7	490	0.653-0.915	0.048-0.091 / 0.140	0.013-0.134 / 0.147	0.194-0.216
pop9 (true)	Anc3–7 + pop7		0.8	0.077 / 0.145	0.037 / 0.145	0.194
pop10	Anc3–7 + pop7	500	0.502-0.690	0.047-0.091 / 0.140	0.021-0.067 / 0.147	0.151-0.167
pop10 (true)	Anc3–7 + pop7		0.6	0.077 / 0.145	0.037 / 0.145	0.150
AdmixedPop2	Mixed1 + Branch3	# rep	α_2	Branch3Loc		
pop8	pop10 + pop2	304	0.782-0.822	0.007-0.040 / 0.040		
	pop10 + Anc1–2	193	0.578-0.756	0.009-0.104 / 0.148		
pop8 (true)	pop10 + pop2		0.8	0.020 / 0.039		

Mixture parameters inferred by *MixMapper* for simulated data, followed by true values for each simulated admixed population. Branch1 and Branch2 are the optimal split points for the mixing populations, with α the proportion of ancestry from Branch1; topologies are shown that that occur for at least 20 of 500 bootstrap replicates. The mixed drift parameters for the three-way admixed pop8 are not well-defined in the simulated tree and are omitted. The branch "Anc3–7" is the common ancestral branch of pops 3–7, and the branch "Anc1–2" is the common ancestral branch of pops 1–2. See Figure 3.2 and the caption of Table 3.2 for descriptions of the parameters and Figure 3.3 for plots of the results.

pop4 is recently admixed between pop3 and pop5, but over a continuous period of 40 generations. Meanwhile, pop8, pop9, and pop10 are all descended from older admixture events, which are similar but with small variations (lower mixture fraction in pop9, 40-generation continuous gene flow in pop10, and subsequent pop2-related admixture into pop8). In the first phase of *MixMapper*, the recently admixed pop4 and pop8 were detected with the 3-population test. From among the other eight populations, a scaffold tree consisting of pop1, pop2, pop3, pop5, pop6, and pop7 provided thorough coverage of the data set and was more additive (max deviation 3.5×10^{-4}) than the secon-best six-population scaffold (5.4×10^{-4}) and the best seven-population scaffold (1.2×10^{-3}). Using this scaffold, *MixMapper* returned very accurate and high-confidence fits for the remaining populations (Figure 3.3E; Table 3.1), with the correct gene flow topologies inferred with 100% confidence for pop4 and pop10, 98% confidence for pop9, and 61% confidence for pop8 (fit as a three-way admixture; 39% of replicates placed the third gene flow source on the branch adjacent to pop2, as shown in Table 3.1). In contrast, *TreeMix* inferred a less accurate admixture model for this data set (Figure 3.3F). *TreeMix* correctly identified pop4 as admixed, and it placed three migration edges among pop7, pop8, pop9, and pop10, but two of the five total admixtures (those originating from the common ancestor of pops 3-5 and the common ancestor of pops 9-10) did not correspond to true events. Also, *TreeMix* did not detect the presence of admixture in pop9 or the pop2-related admixture in pop8.

3.3.2 Application of *MixMapper* to HGDP data

Despite the focus of the HGDP on isolated populations, most of its 53 groups exhibit signs of admixture detectable by the 3-population test, as has been noted previously (Patterson et al., 2012). Thus we hypothesized that applying *MixMapper* to this data set would yield significant insights. Ultimately, we were able to obtain comprehensive results for 20 admixed HGDP populations (Figure 3.4), discussed in detail in the following sections.

3.3.3 Selection of a 10-population unadmixed scaffold tree

To construct an unadmixed scaffold tree for the HGDP data to use in fitting admixtures, we initially filtered the list of 52 populations (having removed San due to ascertainment of our SNP panel in a San individual; see Material and Methods) with the 3-population test, leaving only 20 that are potentially unadmixed. We further excluded Mbuti and Biaka Pygmies, Kalash, Melanesian, and Colombian from the list of candidate populations due to external evidence of admixture (Loh et al., 2013).

It is desirable to include a wide range of populations in the unadmixed scaffold tree to provide both geographic coverage and additional constraints that facilitate the fitting of admixed populations (see Material and Methods). Additionally, incorporating at least four continental groups provides a fairer evaluation of additivity, which is roughly equivalent to measuring discrepancies in fitting phylogenies to quartets of populations. If all populations fall into three or fewer tight clades, however, any quartet must contain at least two populations that are closely related. At the same time, including too many populations can compromise the accuracy of the scaffold. We required that our scaffold tree include representatives of at least four of the five major continental groups in the HGDP data set (Africa, Europe, Oceania, Asia, and the Americas), with at least two populations per group (when available) to clarify the placement of admixing populations and improve the geographical balance. Subject to these conditions, we selected an approximately unadmixed scaffold tree containing 10 populations, which we found to provide a good balance between additivity and comprehensiveness: Yoruba, Mandenka, Papuan, Dai, Lahu, Japanese, Yi, Naxi, Karitiana, and Suruí (Figure 3.4B). These populations constitute the second-most additive (max deviation 1.12×10^{-3}) of 21 similar trees differing only in which East Asian populations are included (range $1.12–1.23 \times 10^{-3}$); we chose them over the most-additive tree because they provide slightly better coverage of Asia. To confirm that modeling these 10 populations as unadmixed in *MixMapper* is sensible, we checked that none of them can be fit in a reasonable way as an admixture on a tree built with the other nine (see Material and Methods). Furthermore, we repeated all of the analyses to follow using nine-population subsets of the unadmixed tree as well as an alternative 11-population tree and confirmed that our results are robust to the choice of scaffold (Figures C.2–C.3; Tables C.1–C.3).

3.3.4 Ancient admixture in the history of present-day European populations

A notable feature of our unadmixed scaffold tree is that it does not contain any European populations. Patterson et al. (2012) previously observed negative f_3 values indicating admix-

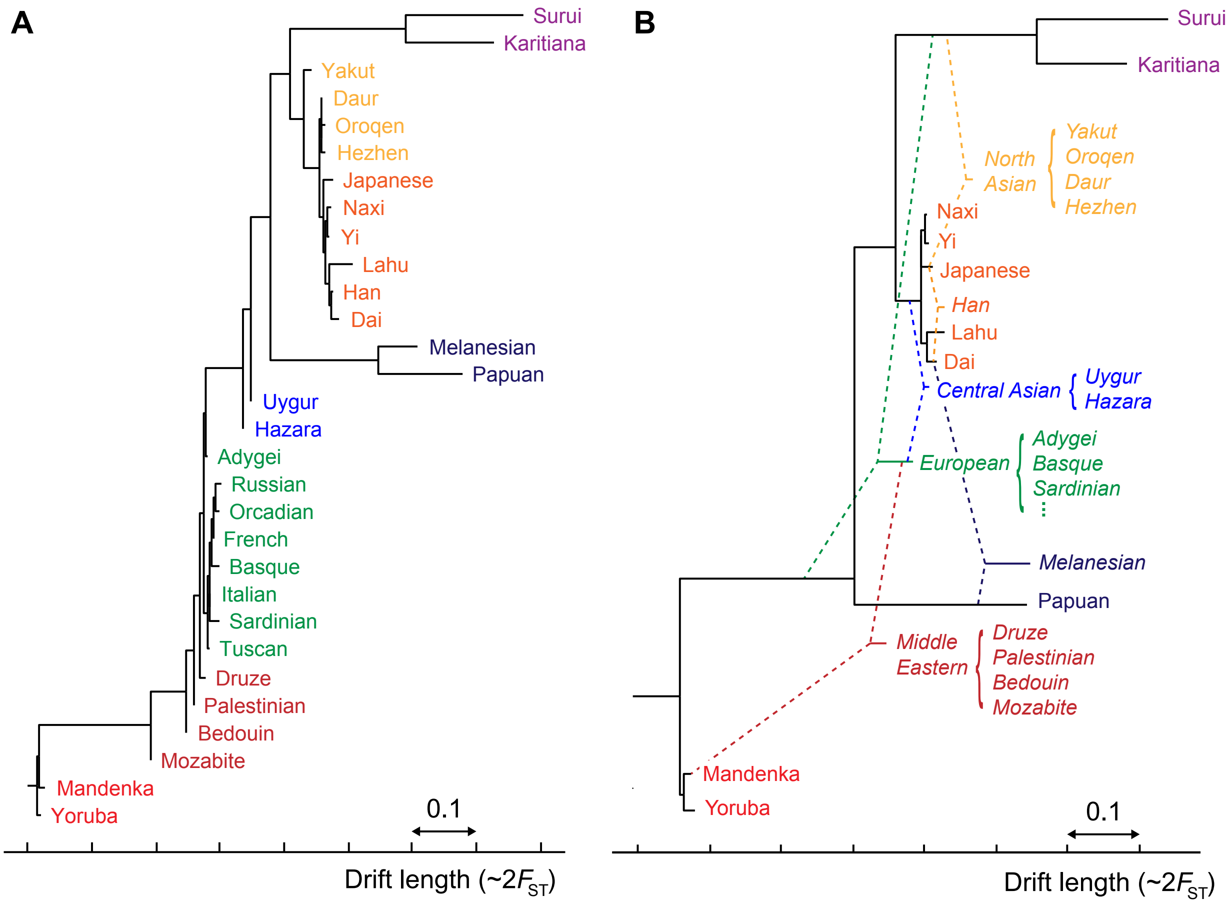


Figure 3.4. Aggregate phylogenetic trees of HGDP populations with and without admixture. (A) A simple neighbor-joining tree on the 30 populations for which *MixMapper* produced high-confidence results. This tree is analogous to the one given by Li et al. (2008b, Figure 1B), and the topology is very similar. (B) Results from *MixMapper*. The populations appear in roughly the same order, but the majority are inferred to be admixed, as represented by dashed lines (cf. Pickrell and Pritchard (2012) and Figure 3.5). Note that drift units are not additive, so branch lengths should be interpreted individually.

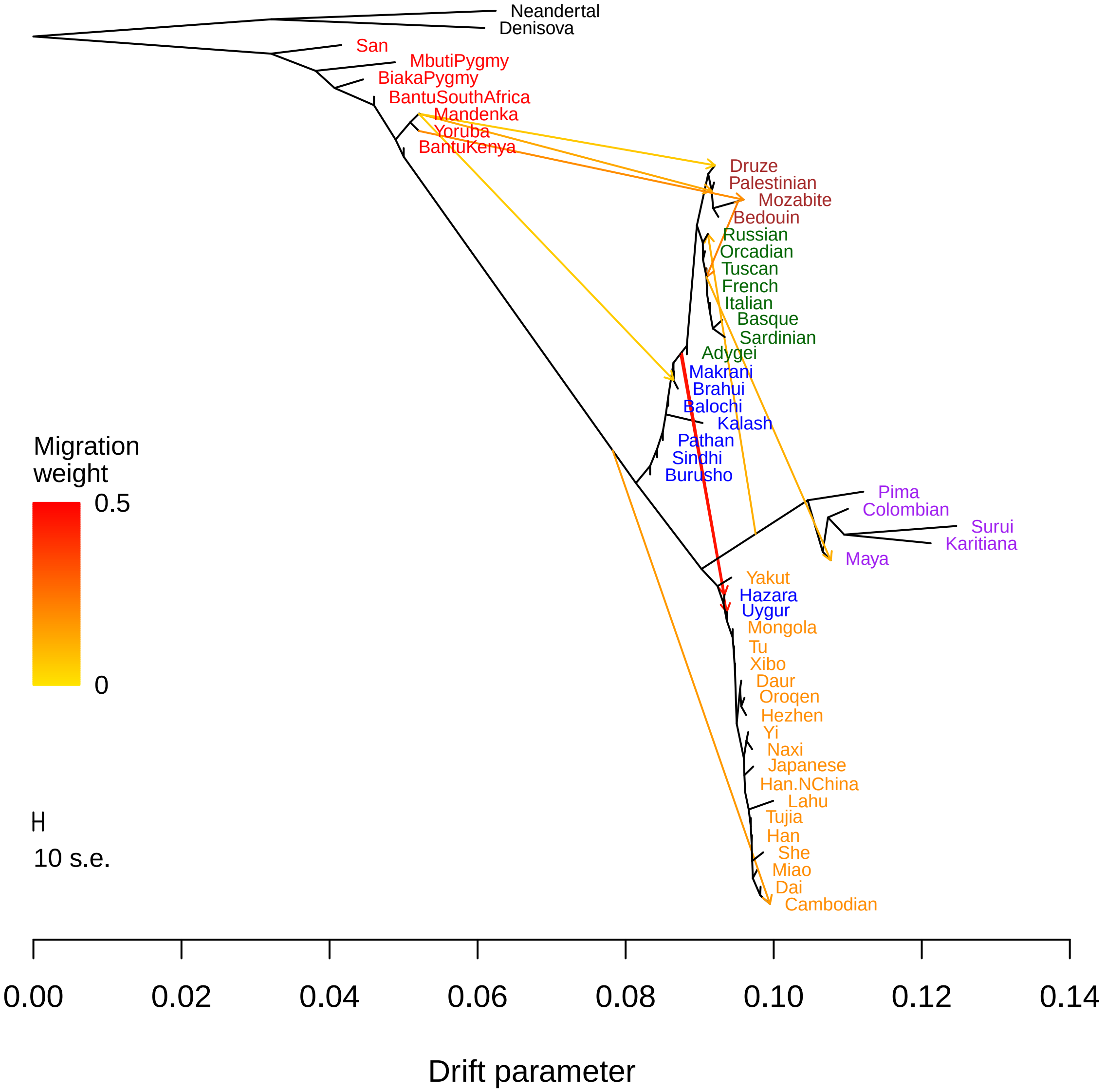


Figure 3.5. *TreeMix* results on the HGDP. Admixture graph for HGDP populations obtained with the *TreeMix* software, as reported in Pickrell and Pritchard (2012). Figure is reproduced from Pickrell and Pritchard (2012) with permission of the authors and under the Creative Commons Attribution License.

Table 3.2. Mixture parameters for Europeans.

AdmixedPop	# rep[a]	α[b]	Branch1Loc (Anc. N-Eur.)[c]	Branch2Loc (Anc. W-Eur.)[c]	MixedDrift[d]
Adygei	500	0.254-0.461	0.033-0.078 / 0.195	0.140-0.174 / 0.231	0.077-0.092
Basque	464	0.160-0.385	0.053-0.143 / 0.196	0.149-0.180 / 0.231	0.105-0.121
French	491	0.184-0.386	0.054-0.130 / 0.195	0.149-0.177 / 0.231	0.089-0.104
Italian	497	0.210-0.415	0.043-0.108 / 0.195	0.137-0.173 / 0.231	0.092-0.109
Orcadian	442	0.156-0.350	0.068-0.164 / 0.195	0.161-0.185 / 0.231	0.096-0.113
Russian	500	0.278-0.486	0.045-0.091 / 0.195	0.146-0.181 / 0.231	0.079-0.095
Sardinian	480	0.150-0.350	0.045-0.121 / 0.195	0.146-0.176 / 0.231	0.107-0.123
Tuscan	489	0.179-0.431	0.039-0.118 / 0.195	0.137-0.177 / 0.231	0.088-0.110

[a] Number of bootstrap replicates (out of 500) placing the mixture between the two branches shown.

[b] Proportion of ancestry from “ancient northern Eurasian” (95% bootstrap confidence interval).

[c] See Figure 3.6A for the definition of the “ancient northern Eurasian” and “ancient western Eurasian” branches in the scaffold tree; Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches (expressed as confidence interval for split point / branch total, as in Figure 3.2A).

[d] Sum of drift lengths $\alpha^2 a + (1-\alpha)^2 b + c$; see Figure 3.2A.

ture in all HGDP Europeans other than Sardinian and Basque. Our *MixMapper* analysis uncovered the additional observation that potential trees containing Sardinian or Basque along with representatives of at least three other continents are noticeably less additive than four-continent trees of the same size without Europeans: from our set of 15 potentially unadmixed populations, none of the 100 most additive 10-population subtrees include Europeans. This points to the presence of admixture in Sardinian and Basque as well as the other European populations.

Using *MixMapper*, we added European populations to the unadmixed scaffold via admixtures (Figure 3.6; Table 3.2). For all eight groups in the HGDP data set, the best fit was as a mixture of a population related to the common ancestor of Karitiana and Suruí (in varying proportions of about 20–40%, with Sardinian and Basque among the lowest and Russian the highest) with a population related to the common ancestor of all non-African populations on the tree. We fit all eight European populations independently, but notably, their ancestors branch from the scaffold tree at very similar points, suggesting a similar broad-scale history. Their branch positions are also qualitatively consistent with previous work that used the 3-population test to deduce ancient admixture for Europeans other than Sardinian and Basque (Patterson et al., 2012). To confirm the signal in Sardinian and Basque, we applied f_4 ratio estimation (Reich et al., 2009; Patterson et al., 2012), which uses allele frequency statistics in a simpler framework to infer mixture proportions. We estimated approximately 20–25% “ancient northern Eurasian” ancestry (Table 3.3), which is in very good agreement with our findings from *MixMapper* (Table 3.2).

At first glance, this inferred admixture might appear improbable on geographical and chronological grounds, but importantly, the two ancestral branch positions do not represent the mixing populations themselves. Rather, there may be substantial drift from the best-fit branch points to the true mixing populations, indicated as branch lengths a and b in Fig-

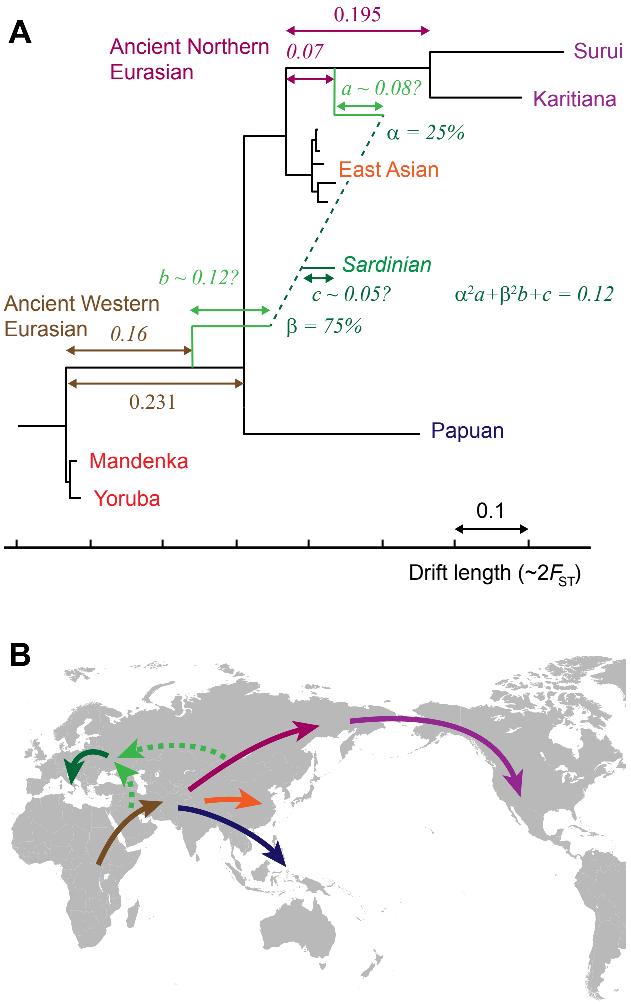


Figure 3.6. Inferred ancient admixture in Europe. (A) Detail of the inferred ancestral admixture for Sardinians (other European populations are similar). One mixing population splits from the unadmixed tree along the common ancestral branch of Native Americans ("Ancient Northern Eurasian") and the other along the common ancestral branch of all non-Africans ("Ancient Western Eurasian"). Median parameter values are shown; 95% bootstrap confidence intervals can be found in Table 3.2. The branch lengths a, b, and c are confounded, so we show a plausible combination. (B) Map showing a sketch of possible directions of movement of ancestral populations. Colored arrows correspond to labeled branches in (A).

Table 3.3. Mixture proportions for Sardinian and Basque from f_4 ratio estimation.

Test pop.	Asian pop.	American pop.	α
Sardinian	Dai	Karitiana	23.3 ± 6.3
Sardinian	Dai	Suruí	24.5 ± 6.7
Sardinian	Lahu	Karitiana	23.1 ± 7.0
Sardinian	Lahu	Suruí	24.7 ± 7.6
Basque	Dai	Karitiana	22.8 ± 7.0
Basque	Dai	Suruí	24.0 ± 7.6
Basque	Lahu	Karitiana	23.1 ± 7.4
Basque	Lahu	Suruí	24.7 ± 8.0

To validate the mixture proportions estimated by *MixMapper* for Sardinian and Basque, we applied f_4 ratio estimation. The fraction α of "ancient northern Eurasian" ancestry was estimated as $\alpha = f_4(\text{Papuan, Asian; Yoruba, European}) \,/\, f_4(\text{Papuan, Asian; Yoruba, American})$, where the European population is Sardinian or Basque, Asian is Dai or Lahu, and American is Karitiana or Suruí. Standard errors are from 500 bootstrap replicates. Note that this calculation assumes the topology of the ancestral mixing populations as inferred by *MixMapper* (Figure 3.6A).

ure 3.6A. Unfortunately, these lengths, along with the post-admixture drift c, appear only in a fixed linear combination in the system of f_2 equations (Appendix C.1), and current methods can only give estimates of this linear combination rather than the individual values (Patterson et al., 2012). One plausible arrangement, however, is shown in Figure 3.6A for the case of Sardinian.

3.3.5 Two-way admixtures outside of Europe

We also found several other populations that fit robustly onto the unadmixed tree using simple two-way admixtures (Table 3.4). All of these can be identified as admixed using the 3-population or 4-population tests (Patterson et al., 2012), but with *MixMapper*, we are able to provide the full set of best-fit parameter values to model them in an admixture tree.

First, we found that four populations from North-Central and Northeast Asia—Daur, Hezhen, Oroqen, and Yakut—are likely descended from admixtures between native North Asian populations and East Asian populations related to Japanese. The first three are estimated to have roughly 10–30% North Asian ancestry, while Yakut has 50–75%. Melanesians fit optimally as a mixture of a Papuan-related population with an East Asian population close to Dai, in a proportion of roughly 80% Papuan-related, similar to previous estimates (Reich et al., 2011; Xu et al., 2012). Finally, we found that Han Chinese have an optimal placement as an approximately equal mixture of two ancestral East Asian populations, one related to modern Dai (likely more southerly) and one related to modern Japanese (likely more northerly), corroborating a previous finding of admixture in Han populations between northern and southern clusters in a large-scale genetic analysis of East Asia (HUGO Pan-Asian SNP Consortium, 2009).

Table 3.4. Mixture parameters for non-European populations modeled as two-way admixtures.

AdmixedPop	Branch1 + Branch2[a]	# rep[b]	α[c]	Branch1Loc[d]	Branch2Loc[d]	MixedDrift[e]
Daur	Anc. N-Eur. + Jap.	350	0.067-0.276	0.008-0.126 / 0.195	0.006-0.013 / 0.016	0.006-0.015
	Suruí + Japanese	112	0.021-0.058	0.008-0.177 / 0.177	0.005-0.010 / 0.015	0.005-0.016
Hezhen	Anc. N-Eur. + Jap.	411	0.068-0.273	0.006-0.113 / 0.195	0.006-0.013 / 0.016	0.005-0.029
Oroqen	Anc. N-Eur. + Jap.	410	0.093-0.333	0.017-0.133 / 0.195	0.005-0.013 / 0.015	0.011-0.030
	Karitiana + Japanese	53	0.025-0.086	0.014-0.136 / 0.136	0.004-0.008 / 0.016	0.008-0.026
Yakut	Anc. N-Eur. + Jap.	481	0.494-0.769	0.005-0.026 / 0.195	0.012-0.016 / 0.016	0.030-0.041
Melanesian	Dai + Papuan	424	0.160-0.260	0.008-0.014 / 0.014	0.165-0.201 / 0.247	0.089-0.114
	Lahu + Papuan	54	0.155-0.255	0.003-0.032 / 0.032	0.167-0.208 / 0.249	0.081-0.114
Han	Dai + Japanese	440	0.349-0.690	0.004-0.014 / 0.014	0.008-0.016 / 0.016	0.002-0.006

[a] Optimal split points for mixing populations.
[b] Number of bootstrap replicates (out of 500) placing the mixture between Branch1 and Branch2; topologies are shown that that occur for at least 50 of 500 replicates.
[c] Proportion of ancestry from Branch1 (95% bootstrap confidence interval).
[d] Points at which mixing populations split from their branches (expressed as confidence interval for split point / branch total, as in Figure 3.2A).
[e] Sum of drift lengths $\alpha^2 a + (1-\alpha)^2 b + c$; see Figure 3.2A.

3.3.6 Recent three-way admixtures involving western Eurasians

Finally, we inferred the branch positions of several populations that are well known to be recently admixed (cf. Patterson et al. (2012); Pickrell and Pritchard (2012)) but for which one ancestral mixing population was itself anciently admixed in a similar way to Europeans. To do so, we applied the capability of *MixMapper* to fit three-way admixtures (Figure 3.2B), using the anciently admixed branch leading to Sardinian as one ancestral source branch. First, we found that Mozabite, Bedouin, Palestinian, and Druze, in decreasing order of African ancestry, are all optimally represented as a mixture between an African population and an admixed western Eurasian population (not necessarily European) related to Sardinian (Table 3.5). We also obtained good fits for Uygur and Hazara as mixtures between a western Eurasian population and a population related to the common ancestor of all East Asians on the tree (Table 3.5).

3.3.7 Estimation of ancestral heterozygosity

Using SNPs ascertained in an outgroup to all of our study populations enables us to compute accurate estimates of the heterozygosity (over a given set of SNPs) throughout an unadmixed tree, including at ancestral nodes (see Material and Methods). This in turn allows us to convert branch lengths from f_2 units to easily interpretable drift lengths (see Appendix C.2).

In Figure 3.7C, we show our estimates for the heterozygosity (averaged over all San-ascertained SNPs used) at the most recent common ancestor (MRCA) of each pair of present-day populations in the tree. Consensus values are given at the nodes of Figure 3.7A. The imputed heterozygosity should be the same for each pair of populations with the same MRCA, and indeed, with the new data set, the agreement is excellent (Figure 3.7C). By contrast, inferences of ancestral heterozygosity are much less accurate using HGDP data

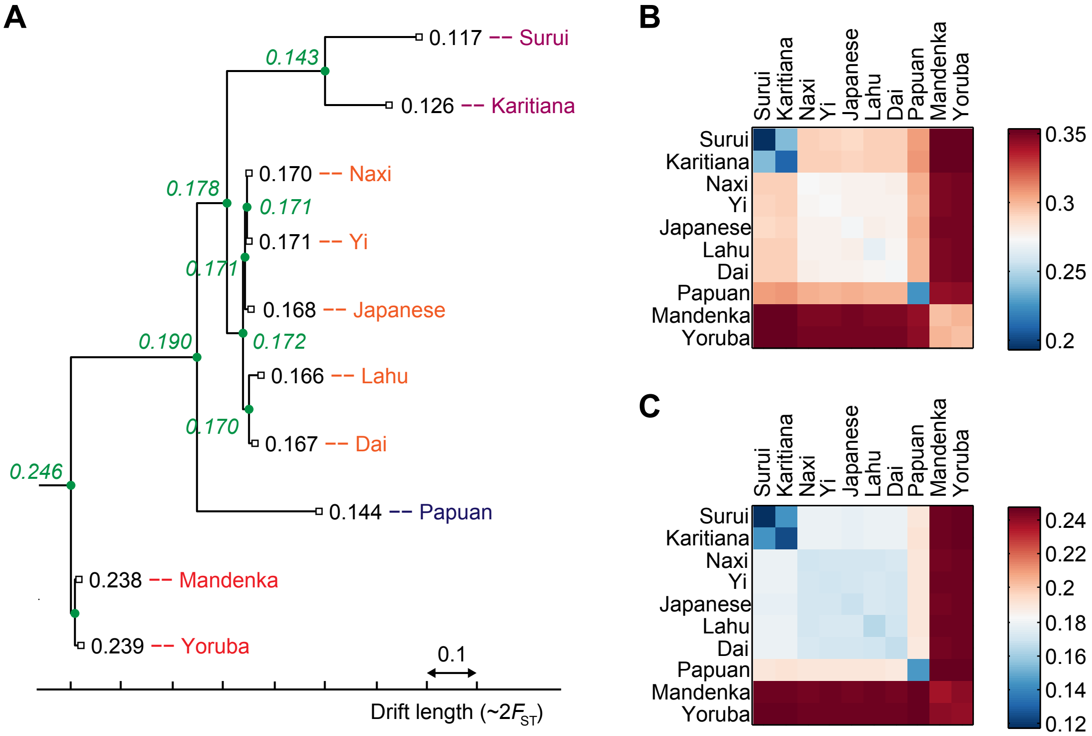


Figure 3.7. Ancestral heterozygosity imputed from original Illumina vs. San-ascertained SNPs. (A) The 10-population unadmixed tree with estimated average heterozygosities using SNPs from Panel 4 (San ascertainment) of the Affymetrix Human Origins array (Patterson et al., 2012). Numbers in black are direct calculations for modern populations, while numbers in green are inferred values at ancestral nodes. (B, C) Computed ancestral heterozygosity at the common ancestor of each pair of modern populations. With unbiased data, values should be equal for pairs having the same common ancestor. (B) Values from a filtered subset of about 250,000 SNPs from the published Illumina array data (Li et al., 2008b). (C) Values from the Human Origins array excluding SNPs in gene regions.

Table 3.5. Mixture parameters for populations modeled as three-way admixtures.

Admixed2	Branch3[a]	# rep[b]	α_2[c]	Branch3Loc[d]	Drift1A[e]	Drift1B[e]	Drift2[e]
Druze	Mandenka	330	0.963-0.988	0.000-0.009 / 0.009	0.081-0.099	0.022-0.030	0.004-0.013
	Yoruba	82	0.965-0.991	0.000-0.010 / 0.010	0.080-0.099	0.022-0.029	0.005-0.013
	Anc. W-Eur.	79	0.881-0.966	0.041-0.158 / 0.232	0.092-0.118	0.000-0.024	0.010-0.031
Palestinian	Anc. W-Eur.	294	0.818-0.901	0.031-0.104 / 0.231	0.093-0.123	0.000-0.021	0.007-0.022
	Mandenka	146	0.909-0.937	0.000-0.009 / 0.009	0.083-0.097	0.022-0.029	0.001-0.007
	Yoruba	53	0.911-0.938	0.000-0.010 / 0.010	0.077-0.098	0.021-0.029	0.001-0.008
Bedouin	Anc. W-Eur.	271	0.767-0.873	0.019-0.086 / 0.231	0.094-0.122	0.000-0.022	0.012-0.031
	Mandenka	176	0.856-0.923	0.000-0.008 / 0.008	0.080-0.099	0.023-0.030	0.006-0.018
Mozabite	Mandenka	254	0.686-0.775	0.000-0.009 / 0.009	0.088-0.109	0.012-0.022	0.017-0.032
	Anc. W-Eur.	142	0.608-0.722	0.002-0.026 / 0.232	0.103-0.122	0.000-0.011	0.018-0.035
	Yoruba	73	0.669-0.767	0.000-0.008 / 0.010	0.086-0.108	0.012-0.023	0.017-0.031
Hazara	Anc. E-Asian[f]	497	0.364-0.471	0.010-0.024 / 0.034	0.080-0.115	0.004-0.034	0.004-0.013
Uygur	Anc. E-Asian[f]	500	0.318-0.438	0.007-0.023 / 0.034	0.088-0.123	0.000-0.027	0.000-0.009

[a] Optimal split point for the third ancestry component. The first two components are represented by a parent population splitting from the (admixed) Sardinian branch.
[b] Number of bootstrap replicates placing the third ancestry component on Branch3; topologies are shown that that occur for at least 50 of 500 replicates.
[c] Proportion of European-related ancestry (95% bootstrap confidence interval).
[d] Point at which mixing population splits from Branch3 (expressed as confidence interval for split point / branch total, as in Figure 3.2A).
[e] Terminal drift parameters; see Figure 3.2B.
[f] Common ancestral branch of the five East Asian populations in the unadmixed tree (Dai, Japanese, Lahu, Naxi, and Yi).

from the original Illumina SNP array (Li et al., 2008b) because of ascertainment bias (Figure 3.7B); f_2 statistics are also affected but to a lesser degree (Figure 3.8), as previously demonstrated (Patterson et al., 2012). We used these heterozygosity estimates to express branch lengths of all of our trees in drift units (Appendix C.2).

3.4 Discussion

3.4.1 Comparison with previous approaches

The *MixMapper* framework generalizes and automates several previous admixture inference tools based on allele frequency moment statistics, incorporating them as special cases. Methods such as the 3-population test for admixture and f_4 ratio estimation (Reich et al., 2009; Patterson et al., 2012) have similar theoretical underpinnings, but *MixMapper* provides more extensive information by analyzing more populations simultaneously and automatically considering different tree topologies and sources of gene flow. For example, negative f_3 values—i.e., 3-population tests indicating admixture—can be expressed in terms of relationships among f_2 distances between populations in an admixture tree. In general, 3-population tests can be somewhat difficult to interpret because the surrogate ancestral populations may not in fact be closely related to the true participants in the admixture, e.g., in the "outgroup case" (Reich et al., 2009; Patterson et al., 2012). The relations among the f_2

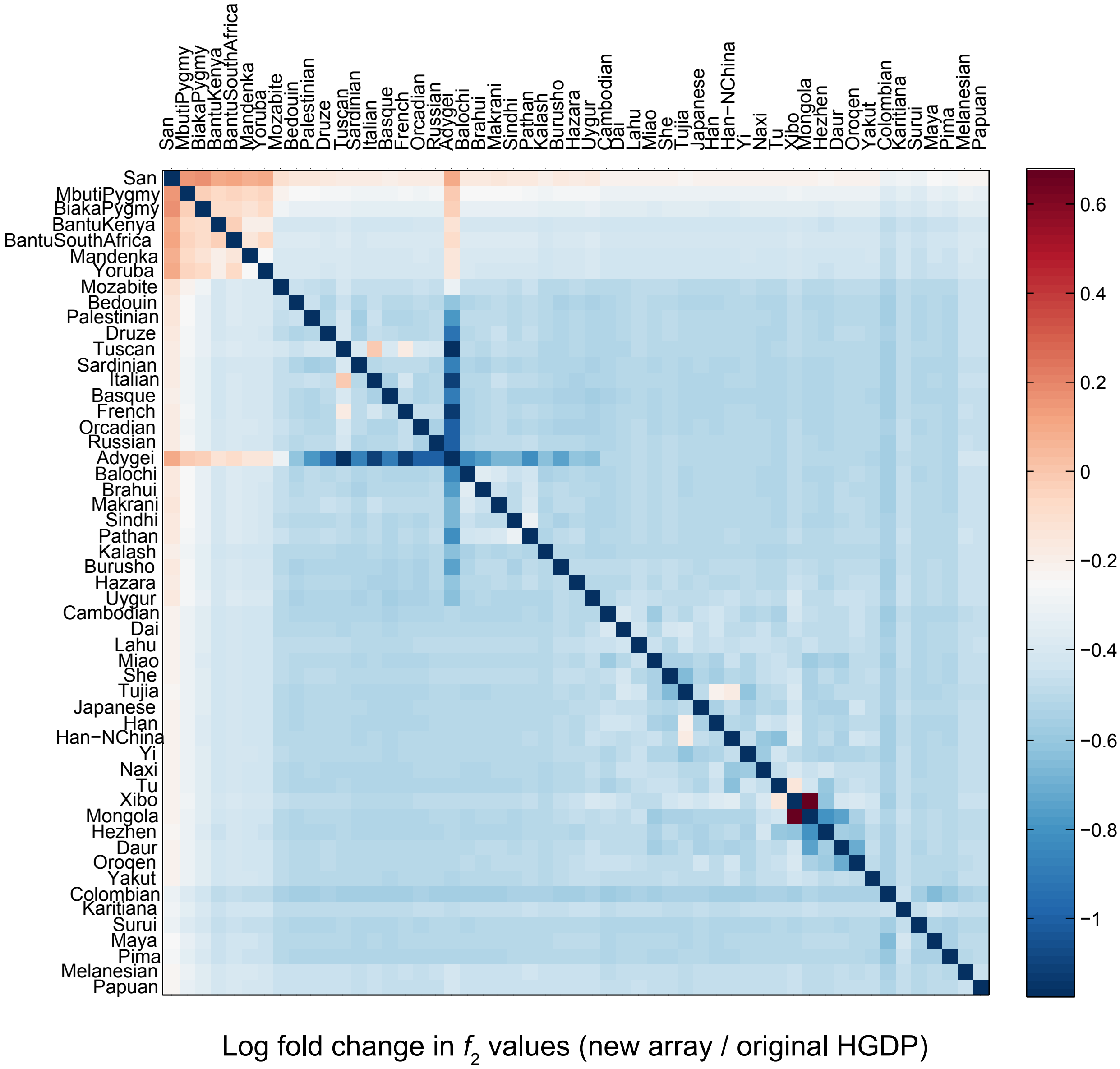


Figure 3.8. Comparison of f_2 distances computed using original Illumina vs. San-ascertained SNPs. The heat map shows the log fold change in f_2 values obtained from the original HGDP data (Li et al., 2008b) versus the San-ascertained data (Patterson et al., 2012) used in this study.

statistics incorporate this situation naturally, however, and solving the full system recovers the true branch points wherever they are. As another example, f_4 ratio estimation infers mixture proportions of a single admixture event from f_4 statistics involving the admixed population and four unadmixed populations situated in a particular topology (Reich et al., 2009; Patterson et al., 2012). Whenever data for five such populations are available, the system of all f_2 equations that *MixMapper* solves to obtain the mixture fraction becomes equivalent to the f_4 ratio computation. More importantly, because *MixMapper* infers all of the topological relationships within an admixture tree automatically by optimizing the solution of the distance equations over all branches, we do not need to specify in advance where the admixture took place—which is not always obvious *a priori*. By using more than five populations, *MixMapper* also benefits from more data points to constrain the fit.

MixMapper also offers significant advantages over the *qpgraph* admixture tree fitting software (Patterson et al., 2012). Most notably, *qpgraph* requires the user to specify the entire topology of the tree, including admixtures, in advance. This requires either prior knowledge of sources of gene flow relative to the reference populations or a potentially lengthy search to test alternative branch locations. *MixMapper* is also faster and provides the capabilities to convert branch lengths into drift units and to perform bootstrap replicates to measure uncertainty in parameter estimates. Furthermore, *MixMapper* is designed to have more flexible and intuitive input and output and better diagnostics for incorrectly specified models. While *qpgraph* does fill a niche of fitting very precise models for small sets of populations, it becomes quite cumbersome for more than about seven or eight, whereas *MixMapper* can be run with significantly larger trees without sacrificing efficiency, ease of use, or accuracy of inferences for populations of interest.

Finally, *MixMapper* differs from *TreeMix* (Pickrell and Pritchard, 2012) in its emphasis on precise and flexible modeling of individual admixed populations. Stylistically, we view *MixMapper* as "semi-automated" as compared to *TreeMix*, which is almost fully automated. Both approaches have benefits: ours allows more manual guidance and lends itself to interactive use, whereas *TreeMix* requires less user intervention, although some care must be taken in choosing the number of gene flow events to include (10 in the HGDP results shown in Figure 3.5) to avoid creating spurious mixtures. With *MixMapper*, we create admixture trees including pre-selected approximately unadmixed populations together with admixed populations of interest, which are added on a case-by-case basis only if they fit reliably as two- or three-way admixtures. In contrast, *TreeMix* returns a single large-scale admixture tree containing all populations in the input data set, which may include some that can be shown to be admixed by other means but are not modeled as such. Thus, these populations might not be placed well on the tree, which in turn could affect the accuracy of the inferred admixture events. Likewise, the populations ultimately modeled as admixed are initially included as part of an unadmixed tree, where (presumably) they do not fit well, which could introduce errors in the starting tree topology that impact the final results.

Indeed, these methodological differences can be seen to affect inferences for both simulated and real data. For our second simulated admixture tree, *MixMapper* very accurately fit the populations with complicated histories (meant to mimic European and Middle Eastern populations), whereas *TreeMix* only recovered portions of the true tree and also added two inaccurate mixtures (Figure 3.3). We believe *TreeMix* was hindered in this case by attempting to fit all of the populations simultaneously and by starting with all of them in an unadmixed tree. In particular, once pop9 (with the lowest proportion of pop7-related admixture) was placed on the unadmixed tree, it likely became difficult to detect as admixed, while pop8's initial placement higher up the tree was likely due to its pop2-related admixture but then obscured this signal in the mixture-fitting phase. Finally, the initial tree shape made populations 3-10 appear to be unequally drifted. Meanwhile, with the HGDP data (Figures 3.4 and 3.5), both methods fit Palestinian, Bedouin, Druze, Mozabite, Uygur, and Hazara as admixed, but *MixMapper* analysis suggested that these populations are better modeled as three-way admixed. *TreeMix* alone fit Brahui, Makrani, Cambodian, and Maya—all of which the 3-population test identifies as admixed but we were unable to place reliably with *MixMapper*—while *MixMapper* alone confidently fit Daur, Hezhen, Oroqen, Yakut, Melanesian, and Han. Perhaps most notably, *MixMapper* alone inferred widespread ancient admixture for Europeans; the closest possible signal of such an event in the *TreeMix* model is a migration edge from an ancestor of Native Americans to Russians. We believe that, as in the simulations, *MixMapper* is better suited to finding a common, ancient admixture signal in a group of populations, and more generally to disentangling complex admixture signals from within a large set of populations, and hence it is able to detect admixture in Europeans when *TreeMix* does not.

To summarize, *MixMapper* offers a suite of features that make it better suited than existing methods for the purpose of inferring accurate admixture parameters in data sets containing many specific populations of interest. Our approach provides a middle ground between *qpgraph*, which is designed to fit small numbers of populations within almost no residual errors, and *TreeMix*, which generates large trees with little manual intervention but may be less precise in complex admixture scenarios. Moreover, *MixMapper*'s speed and interactive design allow the user to evaluate the uncertainty and robustness of results in ways that we have found to be very useful (e.g., by comparing two- vs. three-way admixture models or results obtained using alternative scaffold trees).

3.4.2 Ancient European admixture

Due in part to the flexibility of the *MixMapper* approach, we were able to obtain the notable result that all European populations in the HGDP are best modeled as mixtures between a population related to the common ancestor of Native Americans and a population related to the common ancestor of all non-African populations in our scaffold tree, confirming and extending an admixture signal first reported by Patterson et al. (2012). Our interpretation is that most if not all modern Europeans are descended from at least one large-scale ancient admixture event involving, in some combination, at least one population of Mesolithic European hunter-gatherers; Neolithic farmers, originally from the Near East; and/or other migrants from northern or Central Asia. Either the first or second of these could be related to the "ancient western Eurasian" branch in Figure 3.6, and either the first or third could be related to the "ancient northern Eurasian" branch. Present-day Europeans differ in the amount of drift they have experienced since the admixture and in the proportions of the ancestry components they have inherited, but their overall profiles are similar.

Our results for Europeans are consistent with several previously published lines of evidence (Pinhasi et al., 2012). First, it has long been hypothesized, based on analysis of a few genetic loci (especially on the Y chromosome), that Europeans are descended from ancient admixtures (Semino et al., 2000; Dupanloup et al., 2004; Soares et al., 2010). Our results also suggest an interpretation for a previously unexplained *frappe* analysis of worldwide human population structure (using $K = 4$ clusters) showing that almost all Europeans contain a small fraction of American-related ancestry (Li et al., 2008b). Finally, sequencing of ancient DNA has revealed substantial differentiation in Neolithic Europe between farmers and hunter-gatherers (Bramanti et al., 2009), with the former more closely related to present-day Middle Easterners (Haak et al., 2010) and southern Europeans (Keller et al., 2012; Skoglund et al., 2012) and the latter more similar to northern Europeans (Skoglund et al., 2012), a pattern perhaps reflected in our observed northwest-southeast cline in the proportion of "ancient northern Eurasian" ancestry (Table 3.2). Further analysis of ancient DNA may help shed more light on the sources of ancestry of modern Europeans (Der Sarkissian et al., 2013).

One important new insight of our European analysis is that we detect the same signal of admixture in Sardinian and Basque as in the rest of Europe. As discussed above, unlike other Europeans, Sardinian and Basque cannot be confirmed to be admixed using the 3-population test (as in Patterson et al. (2012)), likely due to a combination of less "ancient northern Eurasian" ancestry and more genetic drift since the admixture (Table 3.2). The first point is further complicated by the fact that we have no unadmixed "ancient western Eurasian" population available to use as a reference; indeed, Sardinians themselves are often taken to be such a reference. However, *MixMapper* uncovered strong evidence for admixture in Sardinian and Basque through additivity-checking in the first phase of the program and automatic topology optimization in the second phase, discovering the correct arrangement of unadmixed populations and enabling admixture parameter inference, which we then verified directly with f_4 ratio estimation. Perhaps the most convincing evidence of the robustness of this finding is that *MixMapper* infers branch points for the ancestral mixing populations that are very similar to those of other Europeans (Table 3.2), a concordance that is most parsimoniously explained by a shared history of ancient admixture among Sardinian, Basque, and other European populations. Finally, we note that because we fit all European populations without assuming Sardinian or Basque to be an unadmixed reference, our estimates of the "ancient northern Eurasian" ancestry proportions in Europeans are larger than those in Patterson et al. (2012) and we believe more accurate than others previously reported (Skoglund et al., 2012).

3.4.3 Future directions

It is worth noting that of the 52 populations (excluding San) in the HGDP data set, there were 22 that we were unable to fit in a reasonable way either on the unadmixed tree or as admixtures. In part, this was because our instantaneous-admixture model is intrinsically limited in its ability to capture complicated population histories. Most areas of the world have surely witnessed ongoing low levels of inter-population migration over time, especially between nearby populations, making it difficult to fit admixture trees to the data. We also found cases where having data from more populations would help the fitting process, for example for three-way admixed populations such as Maya where we do not have a sampled group with a simpler admixture history that could be used to represent two of the three components. Similarly, we found that while Central Asian populations such as Burusho, Pathan, and Sindhi have clear signals of admixture from the 3-population test, their ancestry can likely be traced to several different sources (including sub-Saharan Africa in some instances), making them difficult to fit with *MixMapper*, particularly using the HGDP data. Finally, we have chosen here to disregard admixture with archaic humans, which is known to be a small but noticeable component for most populations in the HGDP (Green et al., 2010; Reich et al., 2010). In the future, it will be interesting to extend *MixMapper* and other admixture tree-fitting methods to incorporate the possibilities of multiple-wave and continuous admixture.

In certain applications, full genome sequences are beginning to replace more limited genotype data sets such as ours, but we believe that our methods and SNP-based inference in general will still be valuable in the future. Despite the improving cost-effectiveness of sequencing, it is still much easier and less expensive to genotype samples using a SNP array, and with over 100,000 loci, the data used in this study provide substantial statistical power. Additionally, sequencing technology is currently more error-prone, which can lead to biases in allele frequency-based statistics (Pool et al., 2010).We expect that *MixMapper* will continue to contribute to an important toolkit of population history inference methods based on SNP allele frequency data.

3.5 Material and Methods

3.5.1 Model assumptions and f-statistics

We assume that all SNPs are neutral, biallelic, and autosomal, and that divergence times are short enough that there are no double mutations at a locus. Thus, allele frequency variation—the signal that we harness—is governed entirely by genetic drift and admixture. We model admixture as a one-time exchange of genetic material: two parent populations mix to form a single descendant population whose allele frequencies are a weighted average of the parents'. This model is of course an oversimplification of true mixture events, but it is flexible enough to serve as a first-order approximation.

Our point-admixture model is amenable to allele frequency moment analyses based on f-statistics (Reich et al., 2009; Patterson et al., 2012). We primarily make use of the statistic $f_2(A, B) := E_S[(p_A - p_B)^2]$, where p_A and p_B are allele frequencies in populations A and B, and E_S denotes the mean over all SNPs. Expected values of f_2 can be written in terms of admixture tree parameters as described in Appendix C.1. Linear combinations of f_2 statistics can also be used to form the quantities $f_3(C; A, B) := E_S[(p_C - p_A)(p_C - p_B)]$ and $f_4(A, B; C, D) := E_S[(p_A - p_B)(p_C - p_D)]$, which form the bases of the 3- and 4-population tests for admixture, respectively. For all of our f-statistic computations, we use previously described unbiased estimators (Reich et al., 2009; Patterson et al., 2012).

3.5.2 Constructing an unadmixed scaffold tree

Our *MixMapper* admixture-tree-building procedure consists of two phases (Figure 3.1), the first of which selects a set of unadmixed populations to use as a scaffold tree. We begin by computing f_3 statistics (Reich et al., 2009; Patterson et al., 2012) for all triples of populations P_1, P_2, P_3 in the data set and removing those populations P_3 with any negative values $f_3(P_3; P_1, P_2)$, which indicate admixture. We then use pairwise f_2 statistics to build neighbor-joining trees on subsets of the remaining populations. In the absence of admixture, f_2 distances are additive along paths on a phylogenetic tree (Appendix C.1; cf. Patterson et al. (2012)), meaning that neighbor-joining should recover a tree with leaf-to-leaf distances that are completely consistent with the pairwise f_2 data (Saitou and Nei, 1987). However, with real data, the putative unadmixed subsets are rarely completely additive, meaning that the fitted neighbor-joining trees have residual errors between the inferred leaf-to-leaf distances and the true f_2 statistics. These deviations from additivity are equivalent to non-zero results from the 4-population test for admixture (Reich et al., 2009; Patterson et al., 2012). We therefore evaluate the quality of each putative unadmixed tree according to its maximum error between fitted and actual pairwise distances: for a tree T having distances d between populations P and Q, the deviation from additivity is defined as $\max\{|d(P, Q) - f_2(P, Q)| : P, Q \in T\}$. *MixMapper* computes this deviation on putatively unadmixed subsets of increasing size, retaining a user-specified number of best subsets of each size in a "beam search" procedure to avoid exponential complexity.

Because of model violations in real data, trees built on smaller subsets are more additive, but they are also less informative; in particular, it is beneficial to include populations from as many continental groups as possible in order to provide more potential branch points for admixture fitting. *MixMapper* provides a ranking of the most additive trees of each size as a guide from which the user chooses a suitable unadmixed scaffold. Once the rank-list of trees has been generated, subject to some constraints (e.g., certain populations required), the user can scan the first several most additive trees for a range of sizes, looking for a balance between coverage and accuracy. This can also be accomplished by checking whether removing a population from a proposed tree results in a substantial additivity benefit; if so, it may be wise to eliminate it. Similarly, if the population removed from the tree can be modeled well as admixed using the remaining portion of the scaffold, this provides evidence that it should not be part of the unadmixed tree. Finally, *MixMapper* adjusts the scaffold tree that the user ultimately selects by re-optimizing its branch lengths (maintaining the topology inferred from neighbor-joining) to minimize the sum of squared errors of all pairwise f_2 distances.

Within the above guidelines, users should choose the scaffold tree most appropriate for their purposes, which may involve other considerations. In addition to additivity and overall size, it is sometimes desirable to select more or fewer populations from certain geographical, linguistic, or other categories. For example, including a population in the scaffold that is actually admixed might not affect the inferences as long as it is not too closely related to the admixed populations being modeled. At the same time, it can be useful to have more populations in the scaffold around the split points for an admixed population of interest in order to obtain finer resolution on the branch positions of the mixing populations. For human data in particular, the unadmixed scaffold is only a modeling device; the populations it contains likely have experienced at least a small amount of mixture. A central goal in building the scaffold is to choose populations such that applying this model will not interfere with the conclusions obtained using the program. The interactive design of *MixMapper* allows the user to tweak the scaffold tree very easily in order to check robustness, and in our analyses, conclusions are qualitatively unchanged for different scaffolds (Figures C.2–C.3; Tables C.1–C.3).

3.5.3 Two-way admixture fitting

The second phase of *MixMapper* begins by attempting to fit additional populations independently as simple two-way admixtures between branches of the unadmixed tree (Figure 3.1). For a given admixed population, assuming for the moment that we know the branches from which the ancestral mixing populations split, we can construct a system of equations of f_2 statistics that allows us to infer parameters of the mixture (Appendix C.1). Specifically, the squared allele frequency divergence $f_2(M, X')$ between the admixed population M and each unadmixed population X' can be expressed as an algebraic combination of known branch lengths along with four unknown mixture parameters: the locations of the split points on the two parental branches, the combined terminal branch length, and the mixture fraction (Figure 3.2A). To solve for the four unknowns, we need at least four unadmixed populations X' that produce a system of four independent constraints on the parameters. This condition is satisfied if and only if the data set contains two populations X'_1 and X'_2 that branch from different points along the lineage connecting the divergence points of the parent populations from the unadmixed tree (Appendix C.1). If the unadmixed tree contains $n > 4$ populations, we obtain a system of n equations in the four unknowns that in theory is dependent. In practice, the equations are in fact slightly inconsistent because of noise in the f_2 statistics and error in the point-admixture model, so we perform least-squares optimization to solve for the unknowns; having more populations helps reduce the impact of noise.

Algorithmically, *MixMapper* performs two-way admixture fitting by iteratively testing each pair of branches of the unadmixed tree as possible sources of the two ancestral mixing populations. For each choice of branches, *MixMapper* builds the implied system of equations and finds the least-squares solution (under the constraints that unknown branch lengths are nonnegative and the mixture fraction α is between 0 and 1), ultimately choosing the pair of branches and mixture parameters producing the smallest residual norm. Our procedure for optimizing each system of equations uses the observation that upon fixing α, the system becomes linear in the remaining three variables (Appendix C.1). Thus, we can optimize the system by performing constrained linear least squares within a basic one-parameter optimization routine over $\alpha \in [0, 1]$. To implement this approach, we applied MATLAB's `lsqlin` and `fminbnd` functions with a few auxiliary tricks to improve computational efficiency (detailed in the code).

3.5.4 Three-way admixture fitting

MixMapper also fits three-way admixtures, i.e., those for which one parent population is itself admixed (Figure 3.2B). Explicitly, after an admixed population M_1 has been added to the tree, *MixMapper* can fit an additional user-specified admixed population M_2 as a mixture between the M_1 terminal branch and another (unknown) branch of the unadmixed tree. The fitting algorithm proceeds in a manner analogous to the two-way mixture case: *MixMapper* iterates through each possible choice of the third branch, optimizing each implied system of equations expressing f_2 distances in terms of mixture parameters. With two admixed populations, there are now $2n+1$ equations, relating observed values of $f_2(M_1, X')$ and $f_2(M_2, X')$ for all unadmixed populations X', and also $f_2(M_1, M_2)$, to eight unknowns: two mixture fractions, α_1 and α_2, and six branch length parameters (Figure 3.2B). Fixing

α_1 and α_2 results in a linear system as before, so we perform the optimization using MATLAB's `lsqlin` within `fminsearch` applied to α_1 and α_2 in tandem. The same mathematical framework could be extended to optimizing the placement of populations with arbitrarily many ancestral admixture events, but for simplicity and to reduce the risk of overfitting, we chose to limit this version of *MixMapper* to three-way admixtures.

3.5.5 Expressing branch lengths in drift units

All of the tree-fitting computations described thus far are performed using pairwise distances in f_2 units, which are mathematically convenient to work with owing to their additivity along a lineage (in the absence of admixture). However, f_2 distances are not directly interpretable in the same way as genetic drift D, which is a simple function of time and population size:

$$D \approx 1 - \exp(-t/2N_e) \approx 2 \cdot F_{ST},$$

where t is the number of generations and N_e is the effective population size (Nei, 1987). To convert f_2 distances to drift units, we apply a new formula, dividing twice the f_2-length of each branch by the heterozygosity value that we infer for the ancestral population at the top of the branch (Appendix C.2). Qualitatively speaking, this conversion corrects for the relative stretching of f_2 branches at different portions of the tree as a function of heterozygosity (Patterson et al., 2012). In order to infer ancestral heterozygosity values accurately, it is critical to use SNPs that are ascertained in an outgroup to the populations involved, which we address further below.

Before inferring heterozygosities at ancestral nodes of the unadmixed tree, we must first determine the location of the root (which is neither specified by neighbor-joining nor involved in the preceding analyses). *MixMapper* does so by iterating through branches of the unadmixed tree, temporarily rooting the tree along each branch, and then checking for consistency of the resulting heterozygosity estimates. Explicitly, for each internal node P, we split its present-day descendants (according to the re-rooted tree) into two groups G_1 and G_2 according to which child branch of P they descend from. For each pair of descendants, one from G_1 and one from G_2, we compute an inferred heterozygosity at P (Appendix C.2). If the tree is rooted properly, these inferred heterozygosities are consistent, but if not, there exist nodes P for which the heterozygosity estimates conflict. *MixMapper* thus infers the location of the root as well as the ancestral heterozygosity at each internal node, after which it applies the drift length conversion as a post-processing step on fitted f_2 branch lengths.

3.5.6 Bootstrapping

In order to measure the statistical significance of our parameter estimates, we compute bootstrap confidence intervals (Efron, 1979; Efron and Tibshirani, 1986) for the inferred branch lengths and mixture fractions. Our bootstrap procedure is designed to account for both the randomness of the drift process at each SNP and the random choice of individuals sampled to represent each population. First, we divide the genome into 50 evenly-sized blocks, with the premise that this scale should easily be larger than that of linkage disequilibrium among our SNPs. Then, for each of 500 replicates, we resample the data set by (a) selecting 50 of these SNP blocks at random with replacement; and (b) for each population group, selecting a random set of individuals with replacement, preserving the number of individuals in the group.

For each replicate, we recalculate all pairwise f_2 distances and present-day heterozygosity values using the resampled SNPs and individuals (adjusting the bias-correction terms to account for the repetition of individuals) and then construct the admixture tree of interest. Even though the mixture parameters we estimate—branch lengths and mixture fractions—depend in complicated ways on many different random variables, we can directly apply the nonparametric bootstrap to obtain confidence intervals (Efron and Tibshirani, 1986). For simplicity, we use a percentile bootstrap; thus, our 95% confidence intervals indicate 2.5 and 97.5 percentiles of the distribution of each parameter among the replicates.

Computationally, we parallelize *MixMapper*'s mixture-fitting over the bootstrap replicates using MATLAB's Parallel Computing Toolbox.

3.5.7 Evaluating fit quality

When interpreting admixture inferences produced by methods such as *MixMapper*, it is important to ensure that best-fit models are in fact accurate. While formal tests for goodness of fit do not generally exist for methods of this class, we use several criteria to evaluate the mixture fits produced by *MixMapper* and distinguish high-confidence results from possible artifacts of overfitting or model violations.

First, we can compare *MixMapper* results to information obtained from other methods, such as the 3-population test (Reich et al., 2009; Patterson et al., 2012). Negative f_3 values indicate robustly that the tested population is admixed, and comparing f_3 statistics for different reference pairs can give useful clues about the ancestral mixing populations. Thus, while the 3-population test relies on similar data to *MixMapper*, its simpler form makes it useful for confirming that *MixMapper* results are reasonable.

Second, the consistency of parameter values over bootstrap replicates gives an indication of the robustness of the admixture fit in question. All results with real data have some amount of associated uncertainty, which is a function of sample sizes, SNP density, intra-population homogeneity, and other aspects of the data. Given these factors, we place less faith in results with unexpectedly large error bars. Most often, this phenomenon is manifested in the placement of ancestral mixing populations: for poorly fitting admixtures, branch choices often change from one replicate to the next, signaling unreliable results.

Third, we find that results where one ancestral population is very closely related to the admixed population and contributes more than 90% of the ancestry are often unreliable. We expect that if we try to fit a non-admixed population as an admixture, *MixMapper* should return a closely related population as the first branch with mixture fraction $\alpha \approx 1$ (and an arbitrary second branch). Indeed, we often observe this pattern in the context of verifying that certain populations make sense to include in the scaffold tree. Further evidence of overfitting comes when the second ancestry component, which contributes only a few percent, either bounces from branch to branch over the replicates, is located at the very tip of a leaf branch, or is historically implausible.

Fourth, for any inferred admixture event, the two mixing populations must be contemporaneous. Since we cannot resolve the three pieces of terminal drift lengths leading to admixed populations (Figure 3.2A) and our branch lengths depend both on population size and absolute time, we cannot say for sure whether this property is satisfied for any given mixture fit. In some cases, however, it is clear that no realization of the variables could possibly be consistent: for example, if we infer an admixture between a very recent branch and a very old one with a small value of the total mixed drift—and hence the terminal drift c—then we can confidently say the mixture is unreasonable.

Finally, when available, we also use prior historical or other external knowledge to guide what we consider to be reasonable. Sometimes, the model that appears to fit the data best has implications that are clearly historically implausible; often when this is true one or more of the evaluation criteria listed above can be invoked as well. Of course, the most interesting findings are often those that are new and surprising, but we subject such results to an extra degree of scrutiny.

3.5.8 Data set and ascertainment

We analyzed a SNP data set from 934 HGDP individuals grouped in 53 populations (Rosenberg et al., 2002; Li et al., 2008b). Unlike most previous studies of the HGDP samples, however, we worked with recently published data generated using the new Affymetrix Axiom Human Origins Array (Patterson et al., 2012), which was designed with a simple ascertainment scheme for accurate population genetic inference (Keinan et al., 2007). It is well known that ascertainment bias can cause errors in estimated divergences among populations (Clark et al., 2005; Albrechtsen et al., 2010), since choosing SNPs based on their properties in modern populations induces non-neutral spectra in related samples. While there do exist methods to correct for ascertainment bias (Nielsen et al., 2004), it is much more desirable to work with *a priori* bias-free data, especially given that typical SNP arrays are designed using opaque ascertainment schemes.

To avoid these pitfalls, we used Panel 4 of the new array, which consists of 163,313 SNPs that were ascertained as heterozygous in the genome of a San individual (Keinan et al., 2007). This panel is special because there is evidence that the San are approximately an outgroup to all other modern-day human populations (Li et al., 2008b; Gronau et al., 2011). Thus, while the Panel 4 ascertainment scheme distorts the San allele frequency spectrum, it is nearly neutral with respect to all other populations. In other words, we can think of the ascertainment as effectively choosing a set of SNPs (biased toward San heterozygosity) at the common ancestor of the remaining 52 populations, after which drift occurs in a bias-free manner. We excluded 61,369 SNPs that are annotated as falling between the transcription start site and end site of a gene in the UCSC Genome Browser database (Fujita et al., 2011). Most of the excluded SNPs are not within actual exons, but as expected, the frequency spectra at these "gene region" loci were slightly shifted toward fixed classes relative to other SNPs, indicative of the action of selection (Figure 3.9). Since we assume neutrality in all of our analyses, we chose to remove these SNPs.

3.5.9 Simulations

Our first simulated tree was generated using the `ms` (Hudson, 2002) command

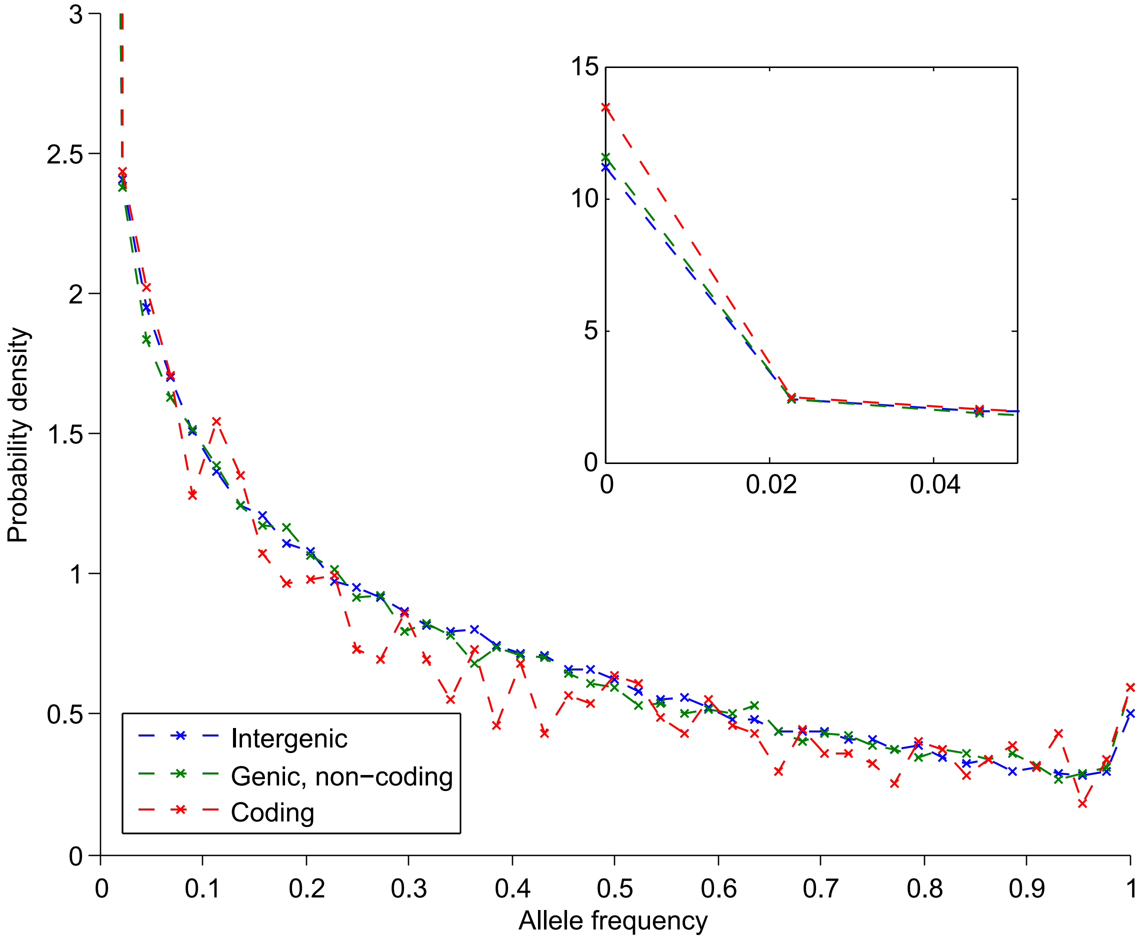


Figure 3.9. Comparison of allele frequency spectra within and outside gene regions. We divided the Panel 4 (San-ascertained) SNPs into three groups: those outside gene regions (101,944), those within gene regions but not in exons (58,110), and those within coding regions (3259). Allele frequency spectra restricted to each group are shown for the Yoruba population. Reduced heterozygosity within exon regions is evident, which suggests the action of purifying selection. (Inset) We observe the same effect in the genic, non-coding spectrum; it is less noticeable but can be seen at the edge of the spectrum.

```
ms 350 500 -t 50 -r 99.9998 500000 -I 7 50 50 50 50 50 50 50 -n 7 2 -n 1 2
-n 2 2 -ej 0.04 2 1 -es 0.02 6 0.4 -ej 0.06 6 3 -ej 0.04 8 5 -ej 0.08 5 4 -ej
0.12 4 3 -ej 0.2 3 1 -ej 0.3 1 7 -en 0.3 7 1.
```

After ascertainment, we used a total of 95,997 SNPs.

Our second simulated tree was generated with the command

```
ms 550 500 -t 50 -r 99.9998 500000 -I 11 50 50 50 50 50 50 50 50 50 50 50 -n
11 2 -n 1 2 -n 2 2 -em 0.002 4 3 253.8 -em 0.004 4 3 0 -es 0.002 8 0.2 -en
0.002 8 2 -ej 0.02 8 2 -ej 0.02 4 5 -ej 0.04 2 1 -ej 0.04 5 3 -es 0.04 12 0.4
-es 0.04 9 0.2 -em 0.042 10 9 253.8 -em 0.044 10 9 0 -ej 0.06 12 7 -ej 0.06
9 7 -ej 0.06 14 10 -ej 0.06 13 10 -ej 0.08 7 6 -ej 0.12 6 3 -ej 0.16 10 3 -ej
0.2 3 1 -ej 0.3 1 11 -en 0.3 11 1.
```

After ascertainment, we used a total of 96,258 SNPs. When analyzing this data set in *TreeMix*, we chose to fit a total of five admixtures based on the residuals of the pairwise distances (maximum of approximately 3 standard errors) and our knowledge that this is the number in the true admixture tree (in order to make for a fair comparison).

3.5.10 Software

Source code for the *MixMapper* software is available at `http://groups.csail.mit.edu/cb/mixmapper/`.

Appendix A

Supporting Information for Compression-Accelerated Search

A.1 CaBLAST

This section describes our compressive implementation of BLAST (Altschul et al., 1990), herein called CaBLAST (Compression-accelerated BLAST). The algorithm is schematically depicted in Figure 1.2.

A.1.1 Preprocessing phase

Conceptually, the preprocessing phase of CaBLAST (Fig. 1.2A) views the original database as an incoming stream of DNA, reading it base by base (with distinct sequences from the database processed in any order) while building four main data structures: the **unique database** and list of **link pointers** with corresponding **edit scripts**—together forming the compressed output of the preprocessing phase—along with a table of **seed 10-mer locations** used internally during preprocessing. As sequence data is processed, it is either:

(a) aligned to existing sequence data in the unique database and represented using a link pointer and edit script; or

(b) added to the unique database (with information about its origin), in which case its 10-mers are added to the table of seeds.

Thus, the unique database roughly contains non-redundant data from the original database, while redundant data is represented using link pointers and edit scripts. For convenience, we refer to the link pointers and edit scripts together as the **links table**.

The main algorithmic challenge in preprocessing is to efficiently identify redundancy within the original database: as we process incoming data, we need to quickly determine whether it aligns well to any part of the unique database being built. We do so by using the seed 10-mer table to rapidly identify seed matches which we then attempt to extend to longer alignments, similar to BLAT (Kent, 2002).

While the overall approach is fairly straightforward, a few important subtleties arise. First, because both matches stored in the links table and sequence data saved to the unique database are likely to consist of sequence fragments rather than complete entries from the original database, it is necessary for the purpose of maintaining search accuracy to lengthen saved fragments so that they overlap (see below for further discussion). Our implementation stores an overlap of 100 bases at transitions between adjacent fragments. The overlap results in an overhead cost per fragment; consequently, we only apply link replacement to alignments of length at least 300 during preprocessing.

Additionally, repetitive genomic elements create anomalously high numbers of matches to a small number of (low-complexity) 10-mer seeds that would substantially reduce performance. Our current implementation circumvents this problem by capping the number of locations stored for a given seed at 500; in practice, aligning regions typically contain at least some seeds that are not of low-complexity and thus the loss of compression is negligible.

We now describe the algorithm more explicitly. The preprocessing phase maintains two pointers that keep track of progress through the incoming data stream:

- a **current base pointer** to the current location being processed; and
- a **last-fragment pointer** to the end of the sequence fragment most recently aligned or copied to the unique database,

and proceeds as follows:

1. Initialize unique database by copying first 10,000 bases of original database; set current base pointer and last-fragment pointer to end of initial region. Initialize table of seed locations with all 10-mers occurring in the copied region. Initialize links table as empty.
2. (Begin loop.) Advance current base pointer one base. Look up 10-mer seed ending at current pointer position, along with its reverse complement, in table of seed locations.
3. For each seed match from table, attempt to extend match (as detailed in next subsection). If extended match of length at least 300 is found:
   (a) Augment links table with link to match and edit script allowing reconstruction (details below).
   (b) Augment unique database with copy of original sequence data preceding alignment that is unaccounted for in unique database: specifically, bases from last-fragment pointer (minus 100 bases) to start of current alignment (plus 100 bases).
   (c) Augment table of seed locations with locations (in the unique database) of all 10-mers occuring in the copied region.
   (d) Move current base pointer and last-fragment pointer to end of alignment and go back to step 2.
4. If 10,000 bases have been processed with no alignments found, copy 10,000-base chunk to unique database, augment table of seed locations accordingly, move both pointers to end of chunk, and go back to step 2. Otherwise, go back directly to step 2.

A.1.2 Details of alignment extension during preprocessing

CaBLAST must rapidly check if a 10-mer seed match extends to an alignment of length 300 or more, allowing only a limited amount of mutation in any chunk of consecutive bases. With this intuition, the algorithm attempts extension in each direction by hopping between matching 4-mers. The extension procedure alternates between the following two steps:

1. Attempt ungapped extension. Scan forward along both sequences looking for the next 4-mer match assuming no gaps. When no 4-mer match is found within 10 bases or when there is less than 50% sequence identity between one 4-mer match and the next, move on to step 2.

2. Attempt gapped extension by local dynamic programming on the next 25 bases of each sequence. Tentatively extend the alignment by taking the path through the dynamic programming table containing the most matching bases. Along this path, identify matching 4-mers and extend the alignment to each successive 4-mer match that occurs within 10 bases of the previous and with at least 50% identity in between (as in step 1). If no acceptable 4-mer match is found, quit. Otherwise, restart step 1 after the last successful extension to a 4-mer match.

The final alignment is accepted as a match if it has length at least 300 and sequence identity exceeding a user-specified threshold (70-90%) in every 100-base window of the alignment.

A.1.3 Edit script compression

Upon identifying a suitable alignment between a sequence fragment and a reference section of the unique database, the sequence data contained in the fragment is compressed by encoding the fragment as a pointer to the original sequence along with a string containing a series of edits that can be applied to the reference segment to produce the aligned fragment. As high sequence identity between the aligned fragments can be assumed from the alignment step, edits can be treated efficiently as islands of difference surrounded by regions of exact matching.

For each differing island, we store a character indicating whether the modification is an insertion (`i`) or substitution/deletion (`s`). We then indicate the number of bases (encoded in octal) since the previous edit, counting along the reference segment, and finally the sequence to insert or substitute. Note here that deletions are treated as substitutions of dashes for the reference bases. Thus, for the following pair of aligned sequences (reference on top, substitutions/deletions in red, insertions in blue for clarity):

```
GTTCACTTATGTATTC--ATATGATTTTGGCAA
GTTCACG--TGTATATTTATATAATTTTGGCAA
```

we generate the following edit script (every other command in green for clarity):

```
s6G--s10ATi2TTs6A
```

We precede each edit script by a 0 or 1, indicating whether the pair of aligned sequences is same-strand or reverse complemented, and also delineating the end of an edit script: on disk, all the edits are concatenated into a single file. The full character set used by the edit scripts thus consists of 16 characters (0–7, A, C, G, T, N, –, i, s) allowing a 4-bit-per-character encoding.

We store pointers in a separate file. Each link pointer contains 20 bytes of information: the start index of the aligned fragment within its originating sequence (4 bytes), the start index of the reference fragment within the unique database (4 bytes), the start index of the edit script within the scripts file (4 bytes), and the lengths of the two aligned sequences (2 bytes each). (Some of this information can be deduced from the edit scripts, but since we build a data structure containing this information for use during search, we included it all in the links table files so as to truly be searching the compressed data without decompressing.)

A.1.4 Choice of parameters

Many constants appear in the above description of our implementation. We did not attempt to optimize all of the various parameters as we intended CaBLAST simply to be a working prototype rather than a recipe for best performance, but the rough rationale and engineering trade-offs involved in each are as follows.

We chose 10,000 bases as the maximum chunk size in order to be large enough to make the impact of link overhead minimal, while small enough to be able to decode hit regions within chunks without an excessive amount of decoding of extraneous regions. The 10,000-base limit also conveniently fit within a 2-byte integer.

We set the overlap of consecutive fragments to 100 bases so that a hit spanning consecutive fragments would likely be picked up as a hit to at least one of them; 100 bases was sufficient for this at typical search sensitivity thresholds for BLAST and BLAT. We then lower-bounded usable fragment lengths by 300 bases to reduce the impact of the overhead cost of overlaps.

As for the alignment extension parameters, all were chosen to strike a reasonable balance between preprocessing speed and gap extension sensitivity. The basic idea is to run dynamic programming on a limited scale in order to detect gaps but give up when further alignment extension seems improbable, so as not to slow down the preprocessing. The parameters reported are simply the ones we tried initially which performed adequately for this purpose.

Finally, the percentage identity requirement on 100-base windows of alignments represents a more significant trade-off between search sensitivity and compression; we discuss this in detail later in this supplement.

A.1.5 Runtime of preprocessing

Algorithmically, the runtime of the CaBLAST preprocessing phase is in between linear and quadratic (with an extremely small constant factor) in the database size depending on the amount of structure in the database. This phase has to be run only once to create the searchable unique database plus links table; there is no need to run it for subsequent searches. The quadratic scaling arises on unstructured sequence data because at each position, the expected number of seed matches found roughly equals the current size of the unique database divided by the number of different seeds—in our 10-mer implementation, close to one million. When the original database contains significant regions of similarity, however, the runtime is closer to linear: in such situations, alignment extension (which takes time only linear in the alignment length) accounts for most of the bases processed; seed matches need only be looked up at the small fraction of positions between alignments. The upshot is that the runtime performance of preprocessing is highly data-dependent.

In practice, we did not take pains to optimize our code for speed and were satisfied being able to complete runs on a standard workstation in a convenient amount of time for testing (a few minutes for 750MB of closely-related microbial sequences and under an hour for 1GB worth of more divergent fly genomes). We are confident that this runtime could be reduced substantially if desired, either by code optimization or by increasing seed lengths to reduce the rate of seed match lookups (at the cost of finding slightly fewer alignments).

A.1.6 Search phase

The search phase of CaBLAST (Fig. 1.2B) applies the procedure below given a query along with coarse and final E-value thresholds:

1. Coarse BLAST. BLAST the query against the unique database to find hits passing the coarse E-value threshold.

2. Link-tracing. For each hit in the unique database, check the links table for other sequence segments outside the unique database that align to the hit region; recover original sequence segments corresponding to the hit by local decompression using stored edit scripts. Extend these segments by 50 bases on each side (in case the linked regions admit longer alignments to the query than the initial hit).

3. Fine BLAST. Re-BLAST against the expanded hits using the final E-value threshold.

A.1.7 Trade-off between compression and accuracy

The percent identity threshold per 100-base window, used during preprocessing to decide whether an alignment qualifies as a link, is of interest because it represents a trade-off between compression and accuracy: relaxing the threshold allows greater database reduction and hence search speedup, but also increases the risk of overlooking alignments during search, due to greater differences between original sequence fragments and their representatives in the unique database. To quantify this trade-off, we ran the CaBLAST preprocessing phase with 100-base identity thresholds ranging from 70-90% on microbial datasets (as discussed in the main text) as well as *Drosophila* subtrees. As expected, accuracy improves while search speedup decreases as the similarity threshold is made stricter (Fig. 1.9). The best parameter choice thus depends on the target application; in our computations we used 80% for the fly genomes and 85% for the microbial datasets.

A.2 CaBLAT

We applied our compressive framework to BLAT in a manner directly analogous to our use of BLAST in the coarse and fine search phases. Specifically, we perform the same preprocessing phase as before to create a unique database and links table.

We then call BLAT directly on the unique database for the coarse search, parse the output (in tab-delimited psl format), perform link-tracing, and regenerate likely hit regions of the original library using the edit scripts from the links table, and finally call BLAT's internal `ffFind()` ("fuzzyfind") and `scoreAli()` routines on these regions to match BLAT's alignment and scoring. Note that BLAT by default uses a simple score threshold based on matches and mismatches between query and target as opposed to BLAST's E-value; thus, we used BLAT's `minIdentity` parameter for coarse and fine search thresholding. In our tests we used the default value of 90 for the final threshold and 80 for the coarse threshold.

A minor technical difference is that BLAT appears to perform slightly better when sequence chunks in the unique database are stored as separate entries, whereas BLAST performs better when all sequences in the unique database are concatenated into one large sequence. We tuned our implementations accordingly. Additionally, BLAT's fuzzyfind algorithm which we used for fine alignment was noticeably more efficient when given hit regions cropped closer to the aligning fragments, so we padded hit regions by only 10 bases during CaBLAT link-tracing.

A.3 Simulated BLAST/BLAT queries

We generated simulated BLAST and BLAT queries by sampling genomic fragments from the genomic library being tested (e.g., for the *melanogaster* subgroup, we took from the *melanogaster* subgroup; for bacteria genera we sampled from the appropriate genus) and then applying random perturbations. More precisely, we randomly selected sequence fragments of random lengths up to 200, after which we replaced a random percentage (0-100%) of the bases in each fragment with random bases. Finally, we rejected a small proportion of these fragments (2% or less), typically from low-complexity regions of the genomes, that aligned to anomalously large numbers of sequence regions and would otherwise have skewed our results.

A.4 Simulated genomes

To simulate clades of recently-diverged species, we used INDELible v1.03 (Fletcher and Yang, 2009), a tool for simulating genome evolution. We used default evolutionary parameters both for the phylogenetic birth-death model and for base-level mutation (JC base substitution probabilities, a geometric indel length distribution, and insertion and deletion rates of 0.08 and 0.12, respectively, relative to the substitution rate).

A.4.1 Fairly testing CaBLAT vs. BLAT

Before performing search, BLAT by default processes the search database indexing k-mers; the time taken by this step can account for a significant amount of BLAT's runtime. For a given query set, we therefore ran BLAT (resp. CaBLAT) on the query set (10,000 simulated queries) followed by a second run on a set containing all queries twice. This does not change the size of the compiled data structure. As initialization time is independent of the number of queries, the difference in times between the two runs gives the time taken for search alone. (To avoid timing discrepancies due to caching of previous files in memory after access, we also preceded this procedure by running the query once first, ignoring the results.) We report the average of five runs.

A.5 Implementation and testing

We implemented CaBLAST and CaBLAT in C++ using the NCBI C++ Toolkit (Version 7.0.0, May 2011) and version 34 of the BLAT source tree, respectively, both for reference runs and in direct calls during the CaBLAST and CaBLAT coarse and fine search phases. Source code is available online at `cast.csail.mit.edu`.

We performed tests on the microbial datasets on a 3.33 GHz Intel Xeon X5680 CPU with 12 MB L3 cache and 48 GB RAM. We ran tests on the fly and simulated datasets on a 3.0 Ghz Intel Xeon CPU with 2 MB L2 cache and 24 GB RAM.

Appendix B

Supporting Information for Admixture Inference Using Linkage Disequilbrium

B.1 Derivations of weighted LD formulas

B.1.1 Expected weighted LD using two diverged reference populations

We now derive equation (2.6) for the expected weighted LD (with respect to random drift) using references A' and B' in place of A and B, retaining the notation of Figure 2.1. Let A' and B' have allele frequencies $p_{A'}(\cdot)$ and $p_{B'}(\cdot)$, and let $\delta'(\cdot) := p_{A'}(\cdot) - p_{B'}(\cdot)$ denote the allele frequency divergences with which we weight the LD $z(x, y)$, giving the two-site statistic

$$a(d) := z(x, y)\delta'(x)\delta'(y).$$

(For brevity, we drop the binning procedure of averaging over SNP pairs (x, y) at distance $|x - y| \approx d$ here.) The value of the random variable $z(x, y)$ is affected by sampling noise as well as genetic drift between A and B, while the random variables $\delta'(x)$ and $\delta'(y)$ are outcomes of genetic drift between A' and B'. These random variables are uncorrelated conditional on the allele frequencies of x and y in A'' and B''. We also assume that x and y are distant enough to have negligible background LD and hence the drifts at the two sites are independent. We then have

$$\begin{aligned} E[a(d)] &= E[z(x, y)\delta'(x)\delta'(y)] \\ &= E[E[z(x, y)\delta'(x)\delta'(y) \mid p_{A''}(x), p_{B''}(x), p_{A''}(y), p_{B''}(y)]] \\ &= E[2\alpha\beta\delta(x)\delta(y)\delta'(x)\delta'(y)e^{-nd}] \\ &= 2\alpha\beta e^{-nd} F_2(A'', B'')^2, \end{aligned}$$

where in the last step the relation $E[\delta(x)\delta'(x)] = E[\delta(y)\delta'(y)] = F_2(A'', B'')$ follows from the fact that the intersection of the drift paths $\delta(\cdot)$ and $\delta'(\cdot)$ is the branch between A'' and B'' (Reich et al., 2009).

B.1.2 Expected weighted LD using one diverged reference population

Using the admixed population C as one reference and a population A' as the other, we have $p_C(\cdot) = \alpha p_A(\cdot) + \beta p_B(\cdot)$ (assuming negligible post-admixture drift), giving weights

$$\delta_{A'C}(\cdot) = p_{A'}(\cdot) - \alpha p_A(\cdot) - \beta p_B(\cdot) = \alpha\delta_{A'A}(\cdot) + \beta\delta_{A'B}(\cdot),$$

where δ_{PQ} denotes the allele frequency difference between populations P and Q. Arguing as above, the expected weighted LD is given by

$$E[a(d)] = E[2\alpha\beta\delta(x)\delta(y)\delta_{A'C}(x)\delta_{A'C}(y)e^{-nd}].$$

To complete the calculation, we compute

$$E[\delta(\cdot)\delta_{A'C}(\cdot)] = \alpha E[\delta(\cdot)\delta_{A'A}(\cdot)] + \beta E[\delta(\cdot)\delta_{A'B}(\cdot)].$$

For the first term, the intersection of the A–B and A'–A drift paths is the A–A'' branch, so $E[\delta(\cdot)\delta_{A'A}(\cdot)] = -F_2(A, A'')$ with the negative sign arising because the paths traverse this branch in opposite directions. For the second term, the intersection of the A–B and A'–B drift paths is the A''–B branch (traversed in the same direction), so $E[\delta(\cdot)\delta_{A'B}(\cdot)] = F_2(B, A'')$. Combining these results gives equation (2.8). (Note that a slight subtlety arises now that we are using population C in our weights: sites x and y can exhibit admixture LD at appreciable distances, so $\delta_{A'C}(x)$ and $\delta_{A'C}(y)$ are not independent. However, only the portions of $\delta_{A'C}(x)$ and $\delta_{A'C}(y)$ arising from post-admixture drift are correlated, and this drift is negligible for typical scenarios we study in which admixture occurred 200 or fewer generations ago.)

B.1.3 Bounding mixture fractions using one reference

We now establish our claim in the main text that the estimator $\hat{\alpha}$ given in equation (2.12) for the mixture fraction α is a lower bound when the reference population A' is diverged from A. Equation (2.12) gives a correct estimate when $A' = A$ but becomes an approximation when there is genetic drift between A and A' or between C and C'. (For accuracy, in this section we relax our usual assumption of negligible drift from C to C'.)

Rearranging equation (2.12), we have by definition

$$\frac{2\hat{\alpha}}{1-\hat{\alpha}} := \frac{\hat{a}_0}{F_2(A', C')^2}. \tag{B.1}$$

From equation (2.7), the amplitude $\hat{a}_0$ is in truth given by

$$\hat{a}_0 = 2\alpha\beta(-\alpha F_2(A, A'') + \beta F_2(B, A''))^2 e^{-n/2N_e},$$

where we have included the post-admixture drift multiplier $e^{-n/2N_e}$ from the C–C' branch.

It follows that

$$\frac{\hat{a}_0}{(-\alpha\beta F_2(A,A'')+\beta^2 F_2(B,A''))^2} = \frac{2\alpha}{\beta}e^{-n/2N_e} < \frac{2\alpha}{1-\alpha}. \tag{B.2}$$

We claim that $F_2(A',C')^2 > (-\alpha\beta F_2(A,A'')+\beta^2 F_2(B,A''))^2$, in which case combining (B.1) and (B.2) gives $\hat{\alpha}/(1-\hat{\alpha}) < \alpha/(1-\alpha)$ and hence $\hat{\alpha} < \alpha$. Indeed, we have

$$\begin{array}{rcl} F_2(A',C') & > & F_2(A'',C) \\ & = & \alpha^2 F_2(A,A'') + \beta^2 F_2(B,A'') \\ & > & -\alpha\beta F_2(A,A'') + \beta^2 F_2(B,A''). \end{array}$$

Squaring both sides appears to give our claim, but we must be careful because it is possible for the final expression to be negative. We will assume A' is closer to A than B, i.e., $F_2(A,A'') < F_2(B,A'')$. Then, if $\alpha < \beta$, the final expression is clearly positive. If $\alpha > \beta$, we have $\alpha^2 F_2(A,A'') > \alpha\beta F_2(A,A'')$ and so

$$F_2(A',C') > \alpha^2 F_2(A,A'') + \beta^2 F_2(B,A'') > \alpha\beta F_2(A,A'') - \beta^2 F_2(B,A'').$$

Thus, squaring the inequality is valid in either case, establishing our bound. From the above we also see that the accuracy of the bound depends on the sizes of the terms that are lost in the approximation—$\alpha F_2(A,A'')$, $F_2(A',A'')$ and $F_2(C,C')$—relative to the term that is kept, $\beta^2 F_2(B,A'')$. In particular, aside from the bound being tighter the closer A' is to A, it is also more useful when the reference A' comes from the minor side $\alpha < 0.5$.

B.1.4 Affine term from population substructure

In the above, we have assumed that population C is homogeneously admixed; i.e., an allele in any random admixed individual from C has a fixed probability α of having ancestry from A and β of having ancestry from B. In practice, many admixed populations experience assortative mating such that subgroups within the population have varying amounts of each ancestry. Heterogeneous admixture among subpopulations creates LD that is independent of genetic distance and not broken down by recombination: intuitively, knowing the value of an allele in one individual changes the prior on the ancestry proportions of that individual, thereby providing information about all other alleles (even those on other chromosomes). This phenomenon causes weighted LD curves to exhibit a nonzero horizontal asymptote, the form of which we now derive.

We model assortative mating by taking α to be a random variable rather than a fixed probability, representing the fact that individuals from different subpopulations of C have different priors on their A ancestry. As before we set $\beta := 1-\alpha$ and we now denote by $\bar{\alpha}$ and $\bar{\beta}$ the population-wide mean ancestry proportions; thus, $\mu_x = \bar{\alpha}p_A(x) + \bar{\beta}p_B(x)$. We wish to compute the expected diploid covariance $E[z(x,y)]$, which we saw in equation (2.2) splits into four terms corresponding to the LD between each copy of the x allele and each copy of the y allele.

Previously, the cross-terms $\operatorname{cov}(X_1,Y_2)$ and $\operatorname{cov}(X_2,Y_1)$ vanished because a homogeneously mixed population does not exhibit inter-chromosome LD. Now, however, writing

$\mathrm{cov}(X_1, Y_2) = E[(X_1 - \mu_x)(Y_2 - \mu_y)]$ as an expectation over individuals from C in the usual way, we find if we condition on the prior α for A ancestry,

$$\begin{aligned}
&E[(X_1 - \mu_x)(Y_2 - \mu_y) \mid p(A \text{ ancestry}) = \alpha] \\
&\quad = \quad E[X_1 - \mu_x \mid p(A \text{ ancestry}) = \alpha] \cdot E[Y_2 - \mu_y \mid p(A \text{ ancestry}) = \alpha] \\
&\quad = \quad (\alpha p_A(x) + \beta p_B(x) - \mu_x)(\alpha p_A(y) + \beta p_B(y) - \mu_y) \\
&\quad = \quad ((\alpha - \bar{\alpha})p_A(x) + (\beta - \bar{\beta})p_B(x))((\alpha - \bar{\alpha})p_A(y) + (\beta - \bar{\beta})p_B(y)) \\
&\quad = \quad ((\alpha - \bar{\alpha})p_A(x) - (\alpha - \bar{\alpha})p_B(x))((\alpha - \bar{\alpha})p_A(y) - (\alpha - \bar{\alpha})p_B(y)) \\
&\quad = \quad (\alpha - \bar{\alpha})^2 \delta(x)\delta(y).
\end{aligned}$$

That is, subpopulations with different amounts of A ancestry make nonzero contributions to the covariance. We can now compute $\mathrm{cov}(X_1, Y_2)$ by taking the expectation of the above over the whole population (i.e., over the random variable α):

$$\mathrm{cov}(X_1, Y_2) = E[(\alpha - \bar{\alpha})^2 \delta(x)\delta(y)] = \mathrm{var}(\alpha)\delta(x)\delta(y) \tag{B.3}$$

and likewise for $\mathrm{cov}(X_2, Y_1)$.

To compute the same-chromosome covariance terms, we split into two cases according to whether or not recombination has occurred between x and y since admixture. In the case that recombination has not occurred—i.e., the ancestry of the chromosomal region between x and y can be traced back as one single chunk to the time of admixture, which occurs with probability e^{-nd}—the region from x to y has ancestry from A with probability α and from B with probability β. Thus,

$$\begin{aligned}
&E[(X_1 - \mu_x)(Y_1 - \mu_y) \mid \text{no recomb}, p(A \text{ ancestry}) = \alpha] \\
&\quad = \quad \alpha E[(X_1 - \mu_x)(Y_1 - \mu_y) \mid A \text{ ancestry}] + \beta E[(X_1 - \mu_x)(Y_1 - \mu_y) \mid B \text{ ancestry}] \\
&\quad = \quad \alpha(p_A(x) - \mu_x)(p_A(y) - \mu_y) + \beta(p_B(x) - \mu_x)(p_B(y) - \mu_y) \\
&\quad = \quad \alpha(\bar{\beta}p_A(x) - \bar{\beta}p_B(x))(\bar{\beta}p_A(y) - \bar{\beta}p_B(y)) + \beta(\bar{\alpha}p_B(x) - \bar{\alpha}p_a(x))(\bar{\alpha}p_B(y) - \bar{\alpha}p_A(y)) \\
&\quad = \quad (\alpha\bar{\beta}^2 + \beta\bar{\alpha}^2)\delta(x)\delta(y).
\end{aligned}$$

Taking the expectation over the whole population,

$$E[(X_1 - \mu_x)(Y_1 - \mu_y) \mid \text{no recomb}] = (\bar{\alpha}\bar{\beta}^2 + \bar{\beta}\bar{\alpha}^2)\delta(x)\delta(y) = \bar{\alpha}\bar{\beta}\delta(x)\delta(y) \tag{B.4}$$

as without assortative mating.

In the case where there has been a recombination, the loci are independent conditioned upon the ancestry proportion α, as in our calculation of the cross-terms; hence,

$$E[(X_1 - \mu_x)(Y_1 - \mu_y) \mid \text{recomb}] = \mathrm{var}(\alpha)\delta(x)\delta(y), \tag{B.5}$$

and this occurs with probability $1 - e^{-nd}$.

Combining equations (B.3), (B.4), and (B.5), we obtain

$$\begin{aligned} E[z(x,y)] &= E[(X-\mu_x)(Y-\mu_y)] \\ &= 2\ \mathrm{var}(\alpha)\delta(x)\delta(y) + 2e^{-nd}\bar{\alpha}\bar{\beta}\delta(x)\delta(y) + 2(1-e^{-nd})\mathrm{var}(\alpha)\delta(x)\delta(y) \\ &= (e^{-nd}(2\bar{\alpha}\bar{\beta} - 2\ \mathrm{var}(\alpha)) + 4\ \mathrm{var}(\alpha))\delta(x)\delta(y). \end{aligned}$$

Importantly, our final expression for $E[z(x,y)]$ still factors as the product of a d-dependent term—now an exponential decay plus a constant—and the allele frequency divergences $\delta(x)\delta(y)$. As it is the product $\delta(x)\delta(y)$ that interacts with our various weighting schemes, the formulas that we have derived for the weighted LD curve $E[a(d)]$—equations (2.4), (2.6), (2.7), and (2.8)—retain the same factors involving F_2 distances and change only in the replacement of $2\alpha\beta e^{-nd}$ with $e^{-nd}(2\bar{\alpha}\bar{\beta} - 2\ \mathrm{var}(\alpha)) + 4\ \mathrm{var}(\alpha)$.

B.2 Testing for admixture

Here we provide details of the weighted LD-based test for admixture we implement in *ALDER*. The test procedure is summarized in the main text; we focus here on technical aspects not given explicitly in Methods.

B.2.1 Determining the extent of LD correlation

The first step of *ALDER* estimates the distance to which LD in the test population is correlated with LD in each reference population. Such correlation suggests shared demographic history that can confound the ALD signal, so it is important to determine the distance to which LD correlation extends and analyze weighted LD curves $\hat{a}(d)$ only for d greater than this threshold. Our procedure is as follows. We successively compute LD correlation for SNP pairs (x,y) within distance bins $d_k < |x-y| < d_{k+1}$, where $d_k = kr$ for some bin resolution r (0.05 cM by default). For each SNP pair (x,y) within a bin, we estimate the LD (i.e., sample covariance between allele counts at x and y) in the test population and the LD in the reference population. We then form the correlation coefficient between the test LD estimates and reference LD estimates over all SNP pairs in the bin. We jackknife over chromosomes to estimate a standard error on the correlation, and we set our threshold after the second bin for which the correlation is insignificant ($p > 0.05$). To reduce dependence on sample size, we then repeat this procedure with successively increasing resolutions up to 0.1 cM and set the final threshold as the maximum of the cutoffs obtained.

B.2.2 Determining significance of a weighted LD curve

To define a formal test for admixture based on weighted LD, we need to estimate the significance of an observed weighted LD curve $\hat{a}(d)$. This question is statistically subtle for several reasons. First, the null distribution of the curve $\hat{a}(d)$ is complex. Clearly the test population C should not be admixed under the null hypothesis, but as we have discussed, shared demography—particularly bottlenecks—can also produce weighted LD. We circumvent this issue by using the pre-tests described in the next section and assume that if the test triple $(C; A', B')$ passes the pre-tests, then under the null hypothesis, non-admixture demographic events have negligible effect on weighted LD beyond the correlation threshold computed above. Even so, the $\hat{a}(d)$ curve still cannot be modeled as random white noise: because SNPs contribute to multiple bins, the curve typically exhibits noticeable autocorrelation. Finally, even if we ignore the issue of colored noise, the question of distinguishing a curve of any type—in our case, an exponential decay—from noise is technically subtle: the difficulty is that a singularity arises in the likelihood surface when the amplitude vanishes, which is precisely the hypothesis that we wish to test (Davies, 1977).

In light of these considerations, we estimate a p-value using the following procedure, which we feel is well-justified despite not being entirely theoretically rigorous. We perform jackknife replicates of the $\hat{a}(d)$ curve computation and fitting, leaving out one chromosome in each replicate, and estimate a standard error for the amplitude and decay constant of the curve using the usual jackknife procedure. We obtain a "z-score" for the amplitude and the decay constant by dividing each by its estimated standard error. Finally, we take the minimum (i.e., less-significant) of these z-scores and convert it to a p-value assuming it comes from a standard normal; we report this p-value as our final significance estimate.

Our intuition for this procedure is that checking the "z-score" of the decay constant essentially tells us whether or not the exponential decay is well-determined: if the $\hat{a}(d)$ curve is actually just noise, then the fitting of jackknife replicates should fluctuate substantially. On the other hand, if the $\hat{a}(d)$ curve has a stable exponential decay constant, then we have good evidence that $\hat{a}(d)$ is actually well-fit by an exponential—and in particular, the amplitude of the exponential is nonzero, meaning we are away from the singularity. In this case the technical difficulty is no longer an issue and the jackknife estimate of the amplitude should in fact give us a good estimate of a z-score that is approximately normal under the null. The "z-score" for the decay constant certainly is not normally distributed—in particular, it is always positive—but taking the minimum of these two scores only makes the test more conservative.

Perhaps most importantly, we have compelling empirical evidence that our z-scores are well-behaved under the null. We applied our test to nine HGDP populations that neither *ALDER* nor the 3-population test identified as admixed; for each test population, we used as references all populations with correlated LD detectable to no more than 0.5 cM. These test triples thus comprise a suite of approximately null tests. We computed Q-Q plots for the reported z-scores and observed that for $z > 0$ (our region of interest), our reported z-scores follow the normal distribution reasonably well, generally erring slightly on the conservative side (Figure B.1). These findings give strong evidence that our significance calculation is sufficiently accurate for practical purposes; in reality, model violation is likely to exert stronger effects than the approximation error in our p-values, and although our empirical tests cannot probe the tail behavior of our statistic, for practical purposes the precise values of p-values less than, say, 10^{-6} are generally inconsequential.

B.2.3 Pre-test thresholds

To ensure that our test is applicable to a given triple $(C; A', B')$, we need to rule out the possibility of demography producing non-admixture-related weighted LD. We do so by computing weighted LD curves for C with weights A'–B', A'–C, and B'–C and fitting an exponential

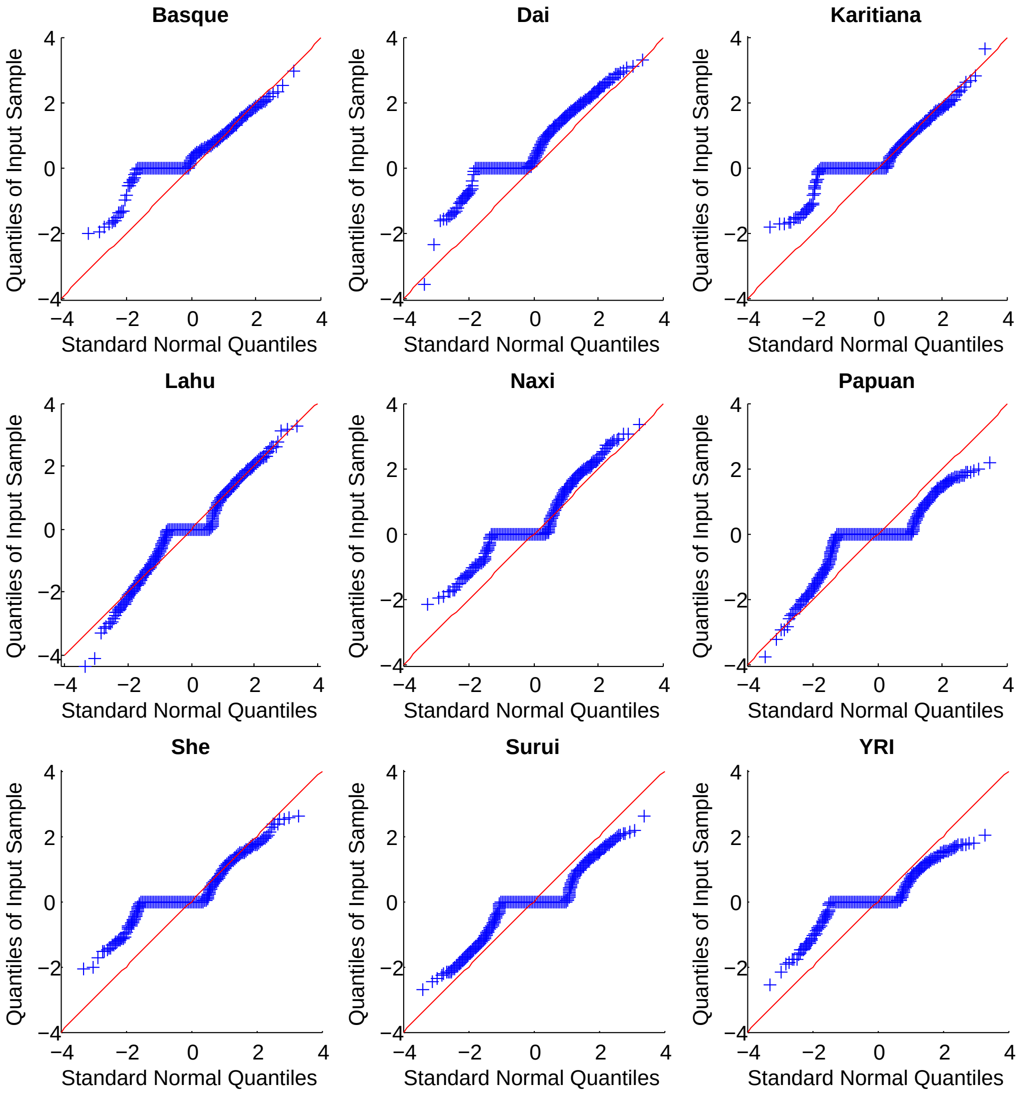


Figure B.1. Q-Q plots comparing *ALDER* z-scores to standard normal on null examples. We show results from nine HGDP populations that neither *ALDER* nor the 3-population test found to be admixed. We are interested in values of $z > 0$; the Q-Q plots show that these values follow the standard normal reasonably well, tending to err on the conservative side.

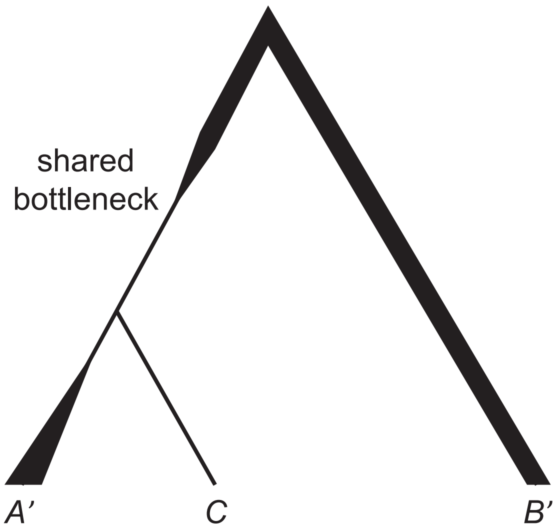


Figure B.2. Non-admixture-related demography producing weighted LD curves. The test population is C and references are A' and B'; the common ancestor of A' and C experienced a recent bottleneck from which C has not yet recovered, leaving long-range LD in C that is potentially correlated to all three possible weighting schemes (A'–B', A'–C, and B'–C).

to each curve. To eliminate the possibility of a shared ancestral bottleneck between C and one of the references, we check that the three estimated amplitudes and decay constants are well-determined; explicitly, we compute a jackknife-based standard error for each parameter and require the implied p-value for the parameter being positive to be less than 0.05. If so, we conclude that whatever LD is present is due to admixture, not other demography, and we report the p-value estimate defined above for the significance of the A'–B' curve as the p-value of our test.

We are aware of one demographic scenario in which the *ALDER* test could potentially return a finding of admixture when the test population is not in fact admixed. As illustrated in Figure B.2, this would occur when A' and C have experienced a shared bottleneck and C has subsequently had a further period of low population size. We do not believe that we have ever encountered such a false positive admixture signal, but to guard against it, we note that if it were to occur, the three decay time constants for the reference pairs A'–B', A'–C, and B'–C would disagree. Thus, along with the test results, *ALDER* returns a warning whenever the three best-fit values of the decay constant do not agree to within 25%.

B.2.4 Multiple-hypothesis correction

In determining statistical significance of test results when testing a population using many pairs of references, we apply a multiple-hypothesis correction that takes into account the number of tests being run. Because some populations in the reference set may be very similar, however, the tests may not be independent. We therefore compute an effective number n_r of distinct references by running PCA on the allele frequency matrix of the reference populations; we take n_r to be the number of singular values required to account for 90% of the total variance. Finally, we apply a Bonferroni correction to the p-values from each test using the effective number $\binom{n_r}{2}$ of reference pairs.

B.3 Coalescent simulations

Here we further validate and explore the properties of weighted LD with entirely *in silico* simulations using the Markovian coalescent simulator MaCS (Chen et al., 2009). These simulations complement the exposition in the main text in which we constructed simulated admixed chromosomes by piecing together haplotype fragments from real HapMap individuals.

B.3.1 Effect of divergence and drift on weighted LD amplitude

To illustrate the effect of using reference populations with varying evolutionary distances from true mixing populations, we performed a set of four simulations in which we varied one reference population in a pair of dimensions: (1) time depth of divergence from the true ancestor, and (2) drift since divergence. In each case, we simulated individuals from three populations A', C', and B', with 22% of C's ancestry derived from a pulse of admixture 40 generations ago from B, where A' and B' diverged 1000 generations ago. We simulated 5 chromosomes of 100 Mb each for 20 diploid individuals from each of A' and B' and 30 individuals from C', with diploid genotypes produced by randomly combining pairs of haploid chromosomes. We assumed an effective population size of 10,000 and set the recombination rate to 10^{-8}. We set the mutation rate parameter to 10^{-9} to have the same effect as using a mutation rate of 10^{-8} and then thinning the data by a factor of 10 (as it would otherwise have produced an unnecessarily large number of SNPs). Finally, we set the MaCS history parameter (the Markovian order of the simulation, i.e., the distance to which the full ancestral recombination graph is maintained) to 10^4 bases.

For the first simulation (Figure 2.2A), we set the divergence of A' and C' to be immediately prior to the gene flow event, altogether resulting in the following MaCS command:

```
macs 140 1e8 -i 5 -h 1e4 -t 0.00004 -r 0.0004 -I 3 40 40 60 -em 0.001 3 2
10000 -em 0.001025 3 2 0 -ej 0.001025 1 3 -ej 0.025 2 3
```

For the second simulation (Figure 2.2B), we increased the drift along the A' terminal branch by reducing the population size by a factor of 20 for the past 40 generations:

```
-en 0 1 0.05 -en 0.001 1 1
```

For the third and fourth simulations (Figure 2.2C,D), we changed the divergence time of A' and C' from 41 to 520 generations, half the distance to the root:

```
-ej 0.001025 1 3   ->   -ej 0.013 1 3
```

We computed weighted LD curves using A'–B' references (Figure 2.2), and the results corroborate our derivation and discussion of equation (2.6). In all cases, the estimated date of admixture is within statistical error of the simulated 40-generation age. The amplitude of the weighted LD curve is unaffected by drift in A' but is substantially reduced by the shorter distance $F_2(A'', B'')$ in the latter two simulations. Increased drift to A' does, however, make the weighted LD curves in the right two panels somewhat noisier than the left two.

B.3.2 Validation of pre-test criteria in test for admixture

To understand the effects of the pre-test criteria stipulated in our LD-based test for admixture, we simulated a variety of population histories with and without mixture. In each case we used the same basic parameter settings as above, except we set the root of each tree to be 4000 generations ago and we simulated 10 chromosomes for each individual instead of 5.

Scenario 1

True admixture 40 generations ago; reference A' diverged 400 generations ago (similar to Figure 2.2C). All pre-tests pass and the our test correctly identifies admixture.

Scenario 2

True admixture 40 generations ago; reference A' diverged 41 generations ago (similar to Figure 2.2A). Because of the proximity of the admixed population C' and the reference A', the test detects long-range correlated LD and concludes that using A' as a reference may produce unreliable results.

Scenario 3

True admixture 40 generations ago; contemporaneous gene flow (of half the magnitude) to the lineage of the reference population A' as well. Again, the pre-test detects long-range correlated LD and concludes that A' is an unsuitable reference.

Scenario 4

No admixture; A and C simply form a clade diverging at half the distance to the root (similar to Figure 2.2C without the gene flow). The test finds no evidence for admixture; weighted LD measurements do not exhibit a decay curve.

Scenario 5

No admixture; A and C diverged 40 generations ago. As above, the test finds no decay in weighted LD. In this scenario the pre-test does detect substantial correlated LD to 1.95 cM because of the proximity of A and C.

Scenario 6

No admixture; same setup as Scenario 4 with addition of recent bottleneck in population C (100-fold reduced population size for the past 40 generations). Here, the test finds no weighted LD decay in the two-reference curve and concludes that there is no evidence for admixture. It does, however, detect decay curves in both one-reference curves (with A–C and B–C weights); these arise because of the strong bottleneck-induced LD within population C.

Scenario 7

No admixture; shared bottleneck: A and C diverged 40 generations ago and their common ancestor underwent a bottleneck of 100-fold reduced population size for the preceding 40 generations. In this case the pre-test detects an enormous amount of correlated LD between A and C and deems A an unsuitable reference.

B.3.3 Sensitivity comparison of 3-population test and LD-based test for admixture

Here we compare the sensitivities of the allele frequency moment-based 3-population test (Reich et al., 2009; Patterson et al., 2012) and our LD-based test for admixture. We simulated a total of 450 admixture scenarios in which we varied three parameters: the age of the branch point A'' (1000, 2000, and 3000 generations), the date n of gene flow (20 to 300 in increments of 20), and the fraction α of A ancestry (50% to 95% in increments of 5%), as depicted in Figure 2.8. In each case we simulated 40 admixed individuals, otherwise using the same parameter settings as in the scenarios above. Explicitly, we ran the commands:

```
macs 160 1e8 -i 10 -h 1e4 -t 0.00004 -r 0.0004 -I 3 40 40 80 -em tMix 3 2
migRate -em tMixStop 3 2 0 -ej tSplit 1 3 -ej 0.1 2 3
```

where `tMix` and `tSplit` correspond to n and the age of A'', while `migRate` and `tMixStop` produce a pulse of gene flow from the B' branch giving C' a fraction α of A ancestry.

We then ran both the 3-population test (f_3) and the *ALDER* test on C' using A' and B' as references (Figure 2.8). The results of these simulations show clearly that the two tests do indeed have complementary parameter ranges of sensitivity. We first observe that the f_3 test is essentially unaffected by the age of admixture (up to the 300 generations we investigate here). As discussed in the main text, its sensitivity is constrained by competition between the admixture signal of magnitude $\alpha\beta F_2(A'', B'')$ and the "off-tree drift" arising from branches off the lineage connecting A' and B' (Reich et al., 2009)—in this case, essentially the quantity $\alpha^2 F_2(A'', C')$. Thus, as the divergence point A'' moves up the lineage, the threshold value of α below which the f_3 test can detect mixture decreases.

The *ALDER* tests behave rather differently, exhibiting a drop-off in sensitivity as the age of admixture increases, with visible noise near the thresholds of sufficient sensitivity. The difference between the f_3 and *ALDER* results is most notable in the bottom panels of Figure 2.8, where the reference A' is substantially diverged from C'. In this case, *ALDER* is still able to identify small amounts of admixture from the B' branch, whereas the f_3 test cannot. Also notable are the vertical swaths of failed tests centered near $\alpha = 0.9, 0.75$, and 0.65 for A'' respectively located at distances 0.75, 0.5, and 0.25 along the branch from the root to A'. This feature of the results arises because the amplitude of the single-reference weighted LD curve with A'–C' weights vanishes near those values of α (see equation (2.8) and Figure 2.3), causing the *ALDER* pre-test to fail. (The two-reference weighted LD exhibits a clear decay curve, but the pre-test is being overly conservative in these cases.) Finally, we also observe that for the smallest choice of mixture age (20 generations), many *ALDER* tests fail. In these cases, the pre-test detects long-range correlated LD with the reference B' and is again overly conservative.

B.3.4 Effect of protracted admixture on weighted LD

The admixture model that we analyze in this manuscript treats admixture as occurring instantaneously in a single pulse of gene flow; however, in real human populations, admixture typically occurs continuously over an extended period of time. Here we explore the effect of protracted admixture on weighted LD curves by simulating scenarios involving continuous migration. We used a setup nearly identical to the simulations above for comparing the f_3 and *ALDER* tests, except here we modified the migration rate and start and end times to correspond to 40% B ancestry that continuously mixed into population C over a period of 0–200 generations ending 40 generations ago. We varied the duration of admixture in increments of 20 generations.

For each simulation, we used *ALDER* to compute the two-reference weighted LD curve and fit an exponential decay. In each case the date of admixture estimated by *ALDER* (Figure 2.7A) falls within the time interval of continuous mixture, as expected (Moorjani et al., 2011). For shorter durations of admixture spanning up to 50 generations or so, the estimated date falls very near the middle of the interval, while it is downward biased for mixtures extending back to hundreds of generations. The amplitude of the fitted exponential also exhibits a downward bias as the mixture duration increases (Figure 2.7B). This behavior occurs because unlike the point admixture case, in which the weighted LD curve follows a simple exponential decay (Figure 2.7C), continuous admixture creates weighted LD that is an average of exponentials with different decay constants (Figure 2.7D).

B.4 FFT computation of weighted LD

In this note we describe how to compute weighted LD (aggregated over distance bins) in time

$$O(m(S + B \log B)),$$

where m is the number of admixed individuals, S is the number of SNPs, and B is the number of bins needed to span the chromosomes. In contrast, the direct method of computing pairwise LD for each individual SNP pair requires $O(mS^2)$ time. In practice our approach offers speedups of over 1000x on typical data sets. We further describe a similar algorithm for computing the single-reference weighted LD polyache statistic that runs in time

$$O(m^2(S + B \log B))$$

with the slight trade-off of ignoring SNPs with missing data.

Our method consists of three key steps: (1) split and factorize the weighted LD product; (2) group factored terms by bin; and (3) apply fast Fourier transform (FFT) convolution. As a special case of this approach, the first two ideas alone allow us to efficiently compute the affine term (i.e., horizontal asymptote) of the weighted LD curve using inter-chromosome SNP pairs.

B.4.1 Two-reference weighted LD

We first establish notation. Say we have an $S \times m$ genotype array $\{c_{x,i}\}$ from an admixed population. Assume for now that there are no missing values, i.e.,

$$c_{x,i} \in \{0, 1, 2\}$$

for x indexing SNPs by position on a genetic map and $i = 1, \ldots, m$ indexing individuals. Given a set of weights w_x, one per SNP, we wish to compute weighted LD of SNP pairs aggregated by inter-SNP distance d:

$$R(d) := \sum_{\substack{|x-y|\approx d \\ x<y}} D_2(x,y) w_x w_y = \frac{1}{2} \sum_{|x-y|\approx d} D_2(x,y) w_x w_y$$

where D_2 is the sample covariance between genotypes at x and y, the diploid analog of the usual LD measure D:

$$\begin{aligned} D_2(x,y) \quad &:= \quad \frac{1}{m-1} \sum_{i=1}^{m} c_{x,i} c_{y,i} - \frac{1}{m(m-1)} \sum_{i=1}^{m} c_{x,i} \sum_{j=1}^{m} c_{y,j} \\ &= \quad \frac{1}{m-1} \sum_{i=1}^{m} c_{x,i} c_{y,i} - \frac{1}{m(m-1)} s_x s_y, \end{aligned} \tag{B.6}$$

where we have defined

$$s_x := \sum_{i=1}^{m} c_{x,i}.$$

Substituting for $D_2(x,y)$, we have

$$\begin{aligned} R(d) \quad &= \quad \frac{1}{2} \sum_{|x-y|\approx d} \left(\frac{1}{m-1} \sum_{i=1}^{m} c_{x,i} c_{y,i} - \frac{1}{m(m-1)} s_x s_y \right) w_x w_y \\ &= \quad \left(\sum_{i=1}^{m} \frac{1}{2(m-1)} \sum_{|x-y|\approx d} c_{x,i} w_x \cdot c_{y,i} w_y \right) - \frac{1}{2m(m-1)} \sum_{|x-y|\approx d} s_x w_x \cdot s_y w_y. \end{aligned} \tag{B.7}$$

We have thus rewritten $R(d)$ as a linear combination of $m + 1$ terms of the form

$$\sum_{|x-y|\approx d} f(x) f(y).$$

(The sum over i consists of m such terms, and the final term accounts for one more.)

In general, sums of the form

$$\sum_{|x-y|\approx d} f(x) g(y)$$

can be efficiently computed by convolution if we first discretize the genetic map on which the SNP positions x and y lie. For notational convenience, choose the distance scale such that a unit distance corresponds to the desired bin resolution. We will compute

$$\sum_{\lfloor x \rfloor - \lfloor y \rfloor = d} f(x)g(y). \tag{B.8}$$

That is, we divide the chromosome into bins of unit distance and aggregate terms $f(x)g(y)$ by the distance between the bin centers of x and y. Note that this procedure does not produce exactly the same result as first subtracting the genetic positions and then binning by $|x-y|$: with our approach, pairs (x, y) that map to a given bin can have actual distances that are off by as much as one full bin width, versus half a bin width with the subtract-then-bin approach. However, we can compensate simply by doubling the bin resolution.

To compute expression (B.8), we write

$$\begin{aligned} \sum_{\lfloor x \rfloor - \lfloor y \rfloor = d} f(x)g(y) &= \sum_{b=0}^{B} \sum_{\lfloor x \rfloor = b} \sum_{\lfloor y \rfloor = b-d} f(x)g(y) \\ &= \sum_{b=0}^{B} \left(\sum_{\lfloor x \rfloor = b} f(x) \right) \left(\sum_{\lfloor y \rfloor = b-d} g(y) \right). \end{aligned} \tag{B.9}$$

Writing

$$F(b) := \sum_{\lfloor x \rfloor = b} f(x), \quad G(b) := \sum_{\lfloor x \rfloor = b} g(x),$$

expression (B.9) becomes

$$\sum_{b=0}^{B} F(b)G(b-d) = (F \star G)(d),$$

a cross-correlation of binned $f(x)$ and $g(y)$ terms.

Computationally, binning f and g to form F and G takes $O(S)$ time, after which the cross-correlation can be performed in $O(B \log B)$ time with a fast Fourier transform. The full computation of the $m+1$ convolutions in equation (B.7) thus takes $O(m(S + B \log B))$ time. In practice we often have $B \log B < S$, in which case the computation is linear in the data size mS.

One additional detail is that we usually want to compute the average rather than the sum of the weighted LD contributions of the SNP pairs in each bin; this requires normalizing by the number of pairs (x, y) that map to each bin, which can be computed in an analogous manner with one more convolution (setting $f \equiv 1$, $g \equiv 1$). Finally, we note that our factorization and binning approach immediately extends to computing weighted LD on inter-chromosome SNP pairs (by putting all SNPs in a chromosome in the same bin), which allows robust estimation of the horizontal asymptote of the weighted LD curve.

Missing Data

The calculations above assumed that the genotype array contained no missing data, but in practice a fraction of the genotype values may be missing. The straightforward non-FFT computation has no difficulty handling missing data, as each pairwise LD term $D_2(x, y)$ can be calculated as a sample covariance over just the individuals successfully genotyped at both x and y. Our algebraic manipulation runs into trouble, however, because if k individuals have a missing value at either x or y, then the sample covariance contains denominators of the form $1/(m-k-1)$ and $1/(m-k)(m-k-1)$—and k varies depending on x and y.

One way to get around this problem is simply to restrict the analysis to sites with no missing values at the cost of slightly reduced power. If a fraction p of the SNPs contain at least one missing value, this workaround reduces the number of SNP pairs available to $(1-p)^2$ of the total, which is probably already acceptable in practice.

We can do better, however: in fact, with a little more algebra (but no additional computational complexity), we can include all pairs of sites (x, y) for which at least one of the SNPs x, y has no missing values, bringing our coverage up to $1-p^2$.

We will need slightly more notation. Adopting `eigenstrat` format, we now let our genotype array consist of values

$$c_{x,i} \in \{0, 1, 2, 9\}$$

where 9 indicates a missing value. (Thus, $\{c_{x,i}\}$ is exactly the data that would be contained in a `.geno` file.) For convenience, we write

$$c_{x,i}^{(0)} := \begin{cases} c_{x,i} & \text{if } c_{x,i} \in \{0,1,2\} \\ 0 & \text{otherwise.} \end{cases}$$

That is, $c_{x,i}^{(0)}$ replaces missing values with 0s. As before we set

$$s_x := \sum_{i : c_{x,i} \neq 9} c_{x,i} = \sum_{i=1}^{m} c_{x,i}^{(0)}$$

to be the sum of all non-missing values at x, which also equals the sum of all $c_{x,i}^{(0)}$ because the missing values have been 0-replaced. Finally, define

$$k_x := \#\{i : c_{x,i} = 9\}$$

to be the number of missing values at site x.

We now wish to compute aggregated weighted LD over pairs (x, y) for which at least one of k_x and k_y is 0. Being careful not to double-count, we have:

$$\begin{aligned}
R(d) &:= \sum_{\substack{|x-y|\approx d \\ x<y \\ k_x=0 \text{ or } k_y=0}} D_2(x,y) w_x w_y \\
&= \frac{1}{2} \sum_{\substack{|x-y|\approx d \\ k_x=0 \text{ and } k_y=0}} D_2(x,y) w_x w_y + \sum_{\substack{|x-y|\approx d \\ k_x=0 \text{ and } k_y\neq 0}} D_2(x,y) w_x w_y \\
&= \sum_{|x-y|\approx d} \frac{I[k_x=0]}{1+I[k_y=0]} D_2(x,y) w_x w_y, \qquad \text{(B.10)}
\end{aligned}$$

where the shorthand $I[\cdot]$ denotes a $\{0,1\}$-indicator.

Now, for a pair of sites (x,y) where x has no missing values and y has k_y missing values,

$$D_2(x,y) = \frac{1}{m-k_y-1} \sum_{i=1}^{m} c_{x,i} c_{y,i}^{(0)} - \frac{1}{(m-k_y)(m-k_y-1)} \left(s_x - \sum_{i=1}^{m} I[c_{y,i}=9] c_{x,i} \right) s_y. \qquad \text{(B.11)}$$

Indeed, we claim the above equation is actually just a rewriting of the standard covariance formula (B.6), appropriately modified now that the covariance is over $m-k_y$ values rather than m:

- In the sum $\sum_{i=1}^{m} c_{x,i} c_{y,i}^{(0)}$, missing values in y have been 0-replaced, so those terms vanish and the sum effectively consists of the desired $m-k_y$ products $c_{x,i}c_{y,i}$.
- Similarly, s_y is equal to the sum of the $m-k_y$ non-missing $c_{y,i}$ values.
- Finally, $s_x - \sum_{i=1}^{m} I[c_{y,i}=9] c_{x,i}$ represents the sum of $c_{x,i}$ over individuals i successfully genotyped at y, written as the sum s_x over all m individuals minus a correction.

Substituting (B.11) into expression (B.10) for $R(d)$ and rearranging, we have

$$\begin{aligned}
R(d) &= \sum_{|x-y|\approx d} \frac{I[k_x=0]}{1+I[k_y=0]} \left(\frac{1}{m-k_y-1} \sum_{i=1}^{m} c_{x,i} c_{y,i}^{(0)} \right. \\
&\qquad \left. - \frac{1}{(m-k_y)(m-k_y-1)} \left(s_x - \sum_{i=1}^{m} I[c_{y,i}=9] c_{x,i} \right) s_y \right) w_x w_y \\
&= \sum_{i=1}^{m} \sum_{|x-y|\approx d} (I[k_x=0] c_{x,i} w_x) \cdot \left(\frac{1}{1+I[k_y=0]} \left(c_{y,i}^{(0)} + \frac{I[c_{y,i}=9] s_y}{m-k_y} \right) \frac{w_y}{m-k_y-1} \right) \\
&\qquad - \sum_{|x-y|\approx d} (I[k_x=0] s_x w_x) \cdot \left(\frac{s_y w_y}{(1+I[k_y=0])(m-k_y)(m-k_y-1)} \right).
\end{aligned}$$

The key point is that we once again have a sum of $m+1$ convolutions, each of the form $\sum_{|x-y|\approx d} f(x)g(y)$, and thus can compute them efficiently as before.

B.4.2 One-reference weighted LD

When computing weighted LD using the admixed population itself as a reference with one other reference population, a polyache statistic must be used to obtain an unbiased estimator (Figure 2.4). The form of the polyache causes complications in our algebraic manipulation; however, if we restrict our attention to SNPs with no missing data, the computation can still be broken into convolutions quite naturally, albeit now requiring $O(m^2)$ FFTs rather than $O(m)$.

As in the two-reference case, the key idea is to split and factorize the weighted LD formula. We treat the terms in the polyache separately and observe that each term takes the form of a constant factor multiplied by a product of sub-terms of the form $S_{r,s}$, $p_A(x)$, or $p_A(y)$. We can use convolution to aggregate the contributions of such a term if we can factor it as a product of two pieces, one depending only on x and the other only on y. Doing so is easy for some terms, namely those that involve only $p_A(x)$, $p_A(y)$, $S_{r,0}$, and $S_{0,s}$, as the latter two sums depend only on x and y, respectively.

The terms involving $S_{r,s}$ with both r and s nonzero are more difficult to deal with but can be written as convolutions by further subdividing them. In fact, we already encountered $S_{1,1} = \sum_{i=1}^{m} c_{x,i}c_{y,i}$ in our two-reference weighted LD computation: the trick there was to split the sum into its m components, one per admixed individual, each of which could then be factored into x-dependent and y-dependent parts and aggregated via convolution.

Exactly the same decomposition works for all of the polyache terms except the one involving $S_{1,1}^2$. For this term, we write

$$S_{1,1}^2 = \sum_{i=1}^{m} c_{x,i}c_{y,i} \sum_{j=1}^{m} c_{x,j}c_{y,j} = \sum_{i=1}^{m}\sum_{j=1}^{m} c_{x,i}c_{x,j} \cdot c_{y,i}c_{y,j},$$

from which we see that splitting the squared sum into m^2 summands allows us to split the x- and y-dependence as desired. The upshot is that at the expense of $O(m^2)$ FFTs (and restricting our analysis to SNPs without missing data), we can also accelerate the one-reference weighted LD computation.

Appendix C

Supporting Information for Admixture Inference Using Moment Statistics

C.1 f-statistics and population admixture

Here we include derivations of the allele frequency divergence equations solved by *MixMapper* to determine the optimal placement of admixed populations. These results were first presented in Reich et al. (2009) and Patterson et al. (2012), and we reproduce them here for completeness, with slightly different emphasis and notation. We also describe in the final paragraph (and in more detail in Material and Methods) how the structure of the equations leads to a particular form of the system for a full admixture tree.

Our basic quantity of interest is the f-statistic f_2, as defined in Reich et al. (2009), which is the squared allele frequency difference between two populations at a biallelic SNP. That is, at SNP locus i, we define

$$f_2^i(A, B) := (p_A - p_B)^2,$$

where p_A is the frequency of one allele in population A and p_B is the frequency of the allele in population B. This is the same as Nei's minimum genetic distance D_{AB} for the case of a biallelic locus (Nei, 1987). As in Reich et al. (2009), we define the unbiased estimator $\hat{f}_2^i(A, B)$, which is a function of finite population samples:

$$\hat{f}_2^i(A, B) := (\hat{p}_A - \hat{p}_B)^2 - \frac{\hat{p}_A(1 - \hat{p}_A)}{n_A - 1} - \frac{\hat{p}_B(1 - \hat{p}_B)}{n_B - 1},$$

where, for each of A and B, $\hat{p}$ is the the empirical allele frequency and n is the total number of sampled alleles.

We can also think of $f_2^i(A, B)$ itself as the outcome of a random process of genetic history. In this context, we define

$$F_2^i(A, B) := E((p_A - p_B)^2),$$

the expectation of $(p_A - p_B)^2$ as a function of population parameters. So, for example, if B is descended from A via one generation of Wright-Fisher genetic drift in a population of size

N, then $F_2^i(A,B) = p_A(1-p_A)/2N$.

While $\hat{f}_2^i(A,B)$ is unbiased, its variance may be large, so in practice, we use the statistic

$$\hat{f}_2(A,B) := \frac{1}{m}\sum_{i=1}^{m} \hat{f}_2^i(A,B),$$

i.e., the average of $\hat{f}_2^i(A,B)$ over a set of m SNPs. As we discuss in more detail in Text C.2, $F_2^i(A,B)$ is not the same for different loci, meaning $\hat{f}_2(A,B)$ will depend on the choice of SNPs. However, we do know that $\hat{f}_2(A,B)$ is an unbiased estimator of the true average $f_2(A,B)$ of $f_2^i(A,B)$ over the set of SNPs.

The utility of the f_2 statistic is due largely to the relative ease of deriving equations for its expectation between populations on an admixture tree. The following derivations are borrowed from (Reich et al., 2009). As above, let the frequency of a SNP i in population X be p_X. Then, for example,

$$\begin{aligned} E(f_2^i(A,B)) &= E((p_A - p_B)^2) \\ &= E((p_A - p_P + p_P - p_B)^2) \\ &= E((p_A - p_P)^2) + E((p_P - p_B)^2) + 2E((p_A - p_P)(p_P - p_B)) \\ &= E(f_2^i(A,P)) + E(f_2^i(B,P)), \end{aligned}$$

since the genetic drifts $p_A - p_P$ and $p_P - p_B$ are uncorrelated and have expectation 0. We can decompose these terms further; if Q is a population along the branch between A and P, then:

$$\begin{aligned} E(f_2^i(A,P)) &= E((p_A - p_P)^2) \\ &= E((p_A - p_Q + p_Q - p_P)^2) \\ &= E((p_A - p_Q)^2) + E((p_Q - p_P)^2) + 2E((p_A - p_Q)(p_Q - p_P)) \\ &= E(f_2^i(A,Q)) + E(f_2^i(Q,P)). \end{aligned}$$

Here, again, $E(p_A - p_Q) = E(p_Q - p_P) = 0$, but $p_A - p_Q$ and $p_Q - p_P$ are not independent; for example, if $p_Q - p_P = -p_P$, i.e. $p_Q = 0$, then necessarily $p_A - p_Q = 0$. However, $p_A - p_Q$ and $p_Q - p_P$ are independent conditional on a single value of p_Q, meaning the conditional expectation of $(p_A - p_Q)(p_Q - p_P)$ is 0. By the double expectation theorem,

$$E((p_A - p_Q)(p_Q - p_P)) = E(E((p_A - p_Q)(p_Q - p_P)|p_Q)) = E(E(0)) = 0.$$

From $E(f_2^i(A,P)) = E(f_2^i(A,Q)) + E(f_2^i(Q,P))$, we can take the average over a set of SNPs to yield, in the notation from above,

$$F_2(A,P) = F_2(A,Q) + F_2(Q,P).$$

We have thus shown that f_2 distances are additive along an unadmixed-drift tree. This property is fundamental for our theoretical results and is also essential for finding admixtures, since, as we will see, additivity does not hold for admixed populations.

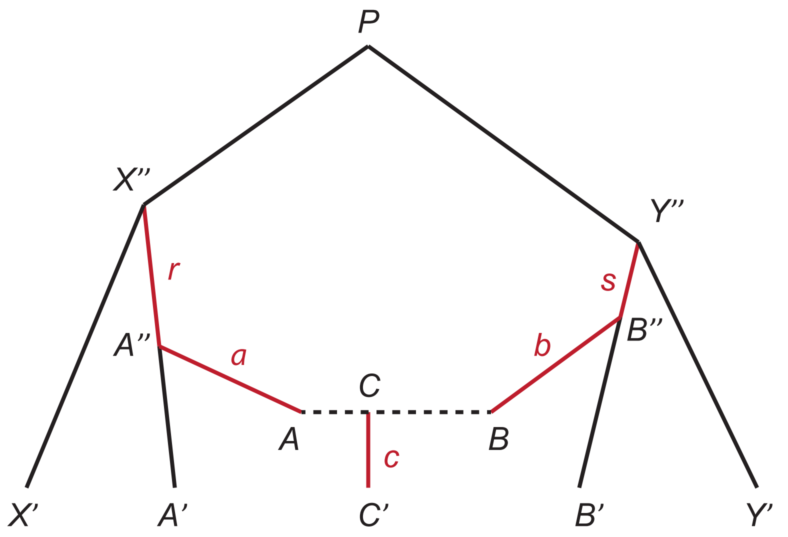


Figure C.1. Schematic of part of an admixture tree. Population C is derived from an admixture of populations A and B with proportion α coming from A. The f_2 distances from C' to the present-day populations A', B', X', Y' give four relations from which we are able to infer four parameters: the mixture fraction α, the locations of the split points A'' and B'' (i.e., r and s), and the combined drift $\alpha^2 a + (1-\alpha)^2 b + c$.

Given a set of populations with allele frequencies at a set of SNPs, we can use the estimator $\hat{f}_2$ to compute f_2 distances between each pair. These distances should be additive if the populations are related as a true tree. Thus, it is natural to build a phylogeny using neighbor-joining (Saitou and Nei, 1987), yielding a fully parameterized tree with all branch lengths inferred. However, in practice, the tree will not exactly be additive, and we may wish to try fitting some population C' as an admixture. To do so, we would have to specify six parameters (in the notation of Figure C.1): the locations on the tree of A'' and B''; the branch lengths $f_2(A'', A)$, $f_2(B'', B)$, and $f_2(C, C')$; and the mixture fraction. These are the variables r, s, a, b, c, and α.

In order to fit C' onto an unadmixed tree (that is, solve for the six mixture parameters), we use the equations for the expectations $F_2(C', Z')$ of the f_2 distances between C' and each other population Z' in the tree. Referring to Figure C.1, with the point admixture model, the allele frequency in C is $p_C = \alpha\, p_A + (1-\alpha)\, p_B$. So, for a single locus, using additivity,

$$\begin{aligned}
E(f_2^i(A', C')) &= E((p_{A'} - p_{C'})^2) \\
&= E((p_{A'} - p_{A''} + p_{A''} - p_C + p_C - p_{C'})^2) \\
&= E((p_{A'} - p_{A''})^2) + E((p_{A''} - \alpha\, p_A - (1-\alpha)\, p_B)^2) + E((p_C - p_{C'})^2) \\
&= E(f_2^i(A', A'')) + \alpha^2 E(f_2^i(A'', A)) \\
&\quad + (1-\alpha)^2 E(f_2^i(A'', B)) + E(f_2^i(C, C')).
\end{aligned}$$

Averaging over SNPs, and replacing $E(f_2(A',C'))$ by the estimator $\hat{f}_2(A',C')$, this becomes

$$\begin{aligned}
\hat{f}_2(A',C') &= F_2(A',X'')-r+\alpha^2 a \\
&\quad +(1-\alpha)^2(r+F_2(X'',Y'')+s+b)+c \\
\implies \hat{f}_2(A',C')-F_2(A',X'') &= (\alpha^2-2\alpha)r+(1-\alpha)^2 s+\alpha^2 a \\
&\quad +(1-\alpha)^2 b+c+(1-\alpha)^2 F_2(X'',Y'').
\end{aligned}$$

The quantities $F_2(X'',Y'')$ and $F_2(A',X'')$ are constants that can be read off of the neighbor-joining tree. Similarly, we have

$$\hat{f}_2(B',C')-F_2(B',Y'')=\alpha^2 r+(\alpha^2-1)s+\alpha^2 a+(1-\alpha)^2 b+c+\alpha^2 F_2(X'',Y'').$$

For the outgroups X' and Y', we have

$$\begin{aligned}
\hat{f}_2(X',C') &= \alpha^2(c+a+r+F_2(X',X'')) \\
&\quad +(1-\alpha)^2(c+b+s+F_2(X'',Y'')+F_2(X',X'')) \\
&\quad +2\alpha(1-\alpha)\,(c+F_2(X',X'')) \\
&= \alpha^2 r+(1-\alpha)^2 s+\alpha^2 a+(1-\alpha)^2 b+c \\
&\quad +(1-\alpha)^2 F_2(X'',Y'')+F_2(X',X'')
\end{aligned}$$

and

$$\hat{f}_2(Y',C')=\alpha^2 r+(1-\alpha)^2 s+\alpha^2 a+(1-\alpha)^2 b+c+\alpha^2 F_2(X'',Y'')+F_2(Y',Y'').$$

Assuming additivity within the neighbor-joining tree, any population descended from A'' will give the same equation (the first type), as will any population descended from B'' (the second type), and any outgroup (the third type, up to a constant and a coefficient of α). Thus, no matter how many populations there are in the unadmixed tree—and assuming there are at least two outgroups X' and Y' such that the points X'' and Y'' are distinct—the system of equations consisting of $E(f_2(P,C'))$ for all P will contain precisely enough information to solve for α, r, s, and the linear combination $\alpha^2 a+(1-\alpha)^2 b+c$. We also note the useful fact that for a fixed value of α, the system is linear in the remaining variables.

C.2 Heterozygosity and drift lengths

One disadvantage to building trees with f_2 statistics is that the values are not in easily interpretable units. For a single locus, the f_2 statistic measures the squared allele frequency change between two populations. However, in practice, one needs to compute an average f_2 value over many loci. Since the amount of drift per generation is proportional to $p(1-p)$, the expected frequency change in a given time interval will be different for loci with different initial frequencies. This means that the estimator $\hat{f}_2$ depends on the distribution of frequencies of the SNPs used to calculate it. For example, within an f_2-based phylogeny, the lengths of non-adjacent edges are not directly comparable.

In order to make use of the properties of f_2 statistics for admixture tree building and still be able to present our final trees in more directly meaningful units, we will show now how f_2 distances can be converted into absolute drift lengths. Again, we consider a biallelic, neutral SNP in two populations, with no further mutations, under a Wright-Fisher model of genetic drift.

Suppose populations A and B are descended independently from a population P, and we have an allele with frequency p in P, $p_A = p + a$ in A, and $p_B = p + b$ in B. The (true) heterozygosities at this locus are $h^i_P = 2p(1-p)$, $h^i_A = 2p_A(1-p_A)$, and $h^i_B = 2p_B(1-p_B)$. As above, we write $\hat{h}^i_A$ for the unbiased single-locus estimator

$$\hat{h}^i_A := \frac{2n_A\hat{p}_A(1-\hat{p}_A)}{n_A - 1},$$

$\hat{h}_A$ for the multi-locus average of $\hat{h}^i_A$, and H^i_A for the expectation of h^i_A under the Wright-Fisher model (and similarly for B and P).

Say A has experienced t_A generations of drift with effective population size N_A since the split from P, and B has experienced t_B generations of drift with effective population size N_B. Then it is well known that $H^i_A = h^i_P(1 - D_A)$, where $D_A = 1 - (1 - 1/(2N_A))^{t_A}$, and $H^i_B = h^i_P(1 - D_B)$. We also have

$$\begin{aligned} H^i_A &= E(2(p+a)(1-p-a)) \\ &= E(h^i_P - 2ap + 2a - 2ap - 2a^2) \\ &= h^i_P - 2E(a^2) \\ &= h^i_P - 2F^i_2(A,P), \end{aligned}$$

so $2F^i_2(A,P) = h^i_P D_A$. Likewise, $2F^i_2(B,P) = h^i_P D_B$ and $2F^i_2(A,B) = h^i_P(D_A + D_B)$. Finally,

$$H^i_A + H^i_B + 2F^i_2(A,B) = h^i_P(1-D_A) + h^i_P(1-D_B) + h^i_P(D_A + D_B) = 2h^i_P.$$

This equation is essentially equivalent to one in Nei (1987), although Nei interprets his version as a way to calculate the expected present-day heterozygosity rather than estimate the ancestral heterozygosity. To our knowledge, the equation has not been applied in the past for this second purpose.

In terms of allele frequencies, the form of h^i_P turns out to be very simple:

$$h^i_P = p_A + p_B - 2p_Ap_B = p_A(1-p_B) + p_B(1-p_A),$$

which is the probability that two alleles, one sampled from A and one from B, are different by state. We can see, therefore, that this probability remains constant in expectation after any amount of drift in A and B. This fact is easily proved directly:

$$E(p_A + p_B - 2p_Ap_B) = 2p - 2p^2 = h^i_P,$$

where we use the independence of drift in A and B.

Let $\hat{h}^i_P := (\hat{h}^i_A + \hat{h}^i_B + 2\hat{f}^i_2(A,B))/2$, and let h_P denote the true average heterozygosity in

P over an entire set of SNPs. Since $\hat{h}_P^i$ is an unbiased estimator of $(h_A^i + h_B^i + 2f_2^i(A,B))/2$, its expectation under the Wright-Fisher model is h_P^i. So, the average $\hat{h}_P$ of $\hat{h}_P^i$ over a set of SNPs is an unbiased (and potentially low-variance) estimator of h_P. If we have already constructed a phylogenetic tree using pairwise f_2 statistics, we can use the inferred branch length $\hat{f}_2(A', P)$ from a present-day population A to an ancestor P in order to estimate $\hat{h}_P$ more directly as $\hat{h}_P = \hat{h}_A + 2\hat{f}_2(A,P)$. This allows us, for example, to estimate heterozygosities at intermediate points along branches or in the ancestors of present-day admixed populations.

The statistic $\hat{h}_P$ is interesting in its own right, as it gives an unbiased estimate of the heterozygosity in the common ancestor of any pair of populations (for a certain subset of the genome). For our purposes, though, it is most useful because we can form the quotient

$$\hat{d}_A := \frac{2\hat{f}_2(A,P)}{\hat{h}_P},$$

where the f_2 statistic is inferred from a tree. This statistic $\hat{d}_A$ is not exactly unbiased, but by the law of large numbers, if we use many SNPs, its expectation is very nearly

$$E(\hat{d}_A) \approx \frac{E(2\hat{f}_2(A,P))}{E(\hat{h}_P)} = \frac{h_P D_A}{h_P} = D_A,$$

where we use the fact that D_A is the same for all loci. Thus $\hat{d}$ is a simple, direct, nearly unbiased moment estimator for the drift length between a population and one of its ancestors. This allows us to convert branch lengths from f_2 distances into absolute drift lengths, one branch at a time, by inferring ancestral heterozygosities and then dividing.

For a terminal admixed branch leading to a present-day population C' with heterozygosity $\hat{h}_{C'}$, we divide twice the inferred mixed drift $c_1 = \alpha^2 a + (1-\alpha)^2 b + c$ (Figure 3.2) by the heterozygosity $\hat{h}_{C'}^* := \hat{h}_{C'} + 2c_1$. This is only an approximate conversion, since it utilizes a common value $\hat{h}_{C'}^*$ for what are really three disjoint branches, but the error should be very small with short drifts.

An alternative definition of $\hat{d}_A$ would be $1 - \hat{h}_A/\hat{h}_P$, which also has expectation (roughly) D_A. In most cases, we prefer to use the definition in the previous paragraph, which allows us to leverage the greater robustness of the f_2 statistics, especially when taken from a multi-population tree.

We note that this estimate of drift lengths is similar in spirit to the widely-used statistic F_{ST}. For example, under proper conditions, the expectation of F_{ST} among populations that have diverged under unadmixed drift is also $1 - (1 - 1/(2N_e))^t$ (Nei, 1987). When F_{ST} is calculated for two populations at a biallelic locus using the formula $(\Pi_D - \Pi_S)/\Pi_D$, where Π_D is the probability two alleles from different populations are different by state and Π_S is the (average) probability two alleles from the same population are different by state (as in Reich et al. (2009) or the measure G'_{ST} in Nei (1987)), then this F_{ST} is exactly half of our $\hat{d}$. As a general rule, drift lengths $\hat{d}$ are approximately twice as large as values of F_{ST} reported elsewhere.

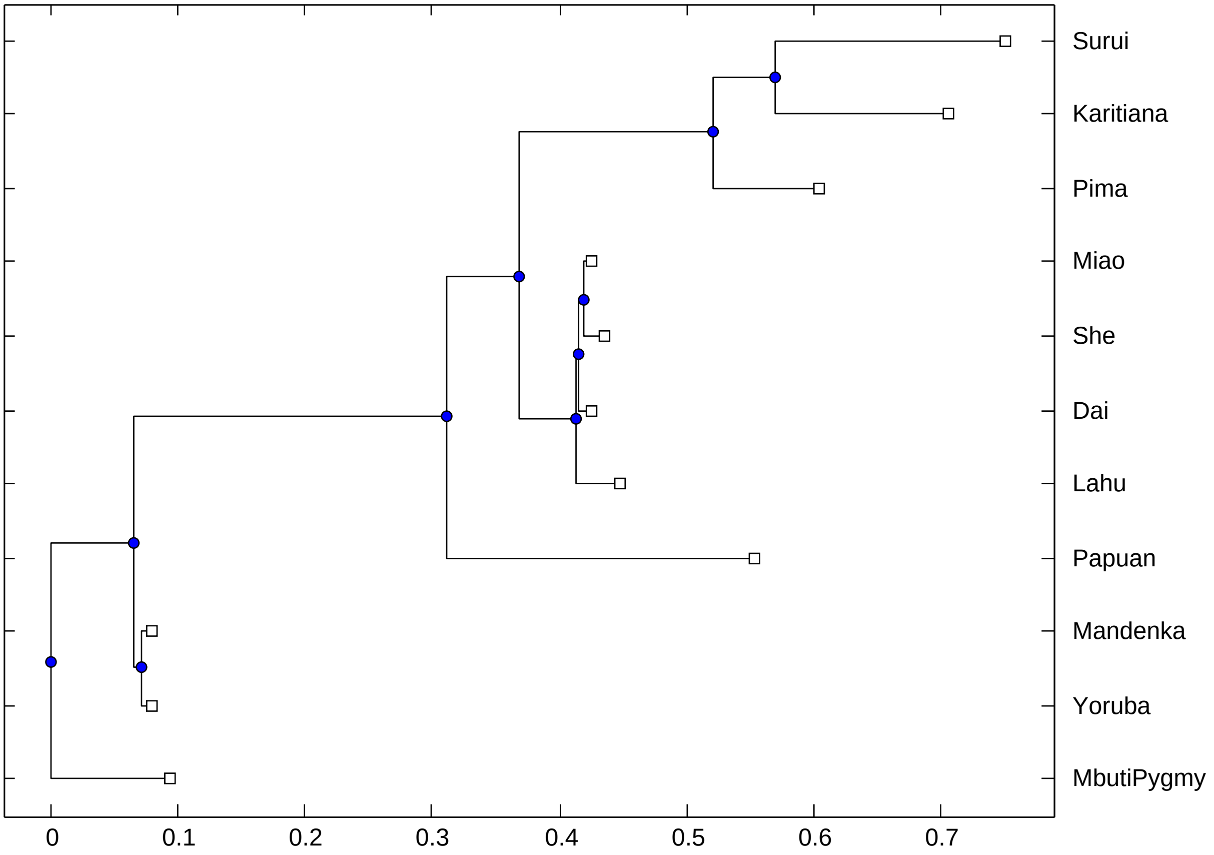


Figure C.2. Alternative scaffold tree with 11 populations used to evaluate robustness of results to scaffold choice. We included Mbuti Pygmy, who are known to be admixed, to help demonstrate that *MixMapper* inferences are robust to deviations from additivity in the scaffold; see Tables C.1–C.3 for full results. Distances are in drift units.

C.3 Robustness of MixMapper HGDP results to scaffold choice

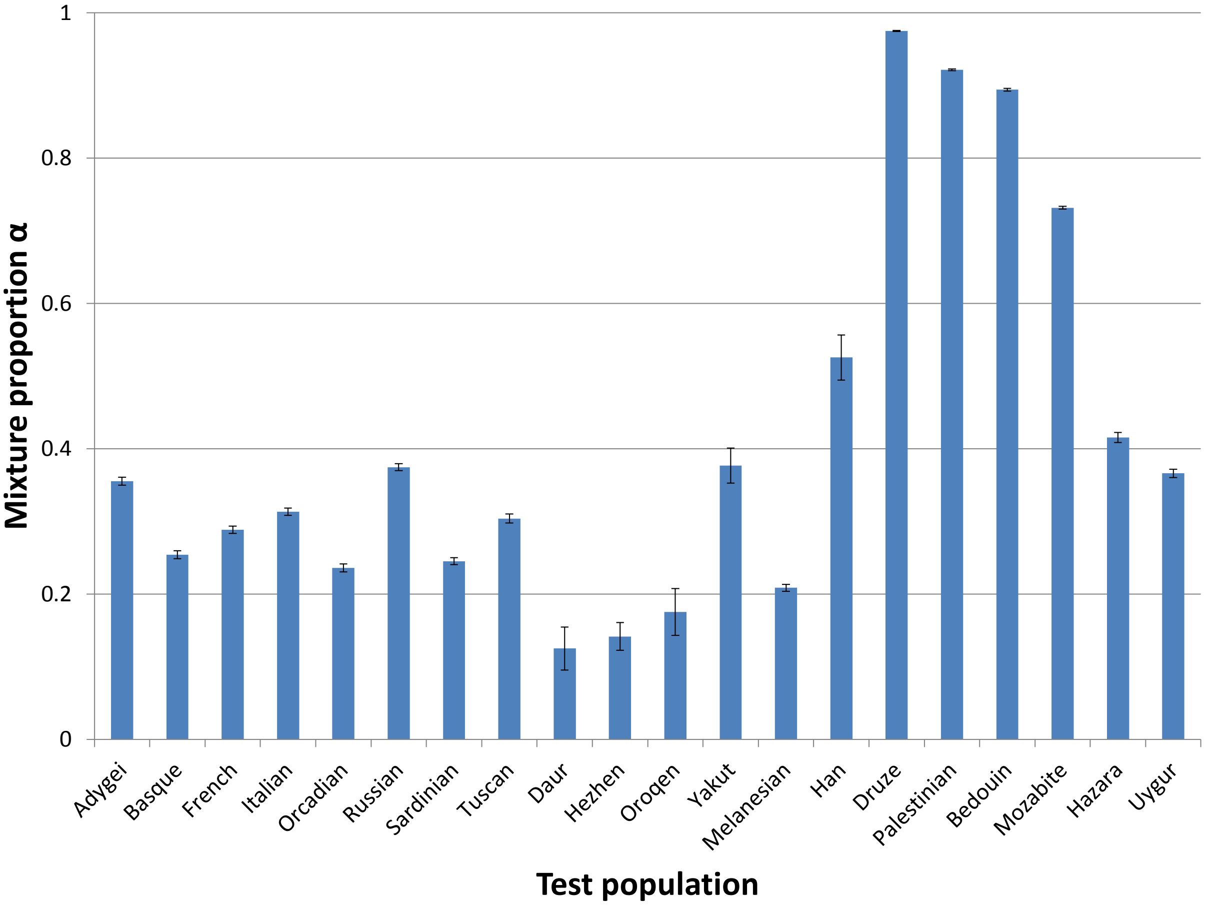


Figure C.3. Summary of mixture proportions α inferred with alternative 9-population scaffold trees. We ran *MixMapper* for all 20 admixed test populations using nine different scaffold trees obtained by removing each population except Papuan one at a time from our full 10-population scaffold. (Papuan is needed to maintain continental representation.) For each test population and each scaffold, we recorded the median bootstrap-inferred value of α over all replicates having branching patterns similar to the primary topology. Shown here are the means and standard deviations of the nine medians. In all cases, α refers to the proportion of ancestry from the first branch as in Tables 3.2–3.5.

Table C.1. Mixture parameters for Europeans inferred with an alternative scaffold tree.

AdmixedPop	# rep^a	α^b	Branch1Loc (Anc. N-Eur.)^c	Branch2Loc (Anc. W-Eur.)^c	MixedDrift^d
Adygei	488	0.278-0.475	0.035-0.078 / 0.151	0.158-0.191 / 0.246	0.078-0.093
Basque	273	0.221-0.399	0.055-0.111 / 0.153	0.164-0.194 / 0.244	0.108-0.124
French	380	0.240-0.410	0.054-0.108 / 0.152	0.165-0.192 / 0.245	0.093-0.106
Italian	427	0.245-0.426	0.047-0.103 / 0.152	0.155-0.188 / 0.246	0.095-0.110
Orcadian	226	0.214-0.387	0.061-0.131 / 0.153	0.174-0.197 / 0.244	0.098-0.116
Russian	472	0.296-0.490	0.047-0.093 / 0.151	0.165-0.197 / 0.246	0.080-0.095
Sardinian	390	0.189-0.373	0.045-0.104 / 0.152	0.160-0.190 / 0.245	0.110-0.125
Tuscan	413	0.238-0.451	0.039-0.096 / 0.152	0.153-0.191 / 0.245	0.093-0.111

Mixture parameters inferred by *MixMapper* for modern-day European populations using an alternative unadmixed scaffold tree containing 11 populations: Yoruba, Mandenka, Mbuti Pygmy, Papuan, Dai, Lahu, Miao, She, Karitiana, Suruí, and Pima (see Figure C.2). The parameter estimates are very similar to those obtained with the original scaffold tree (Table 3.2), with α slightly higher on average. The bootstrap support for the branching position of "ancient northern Eurasian" plus "ancient western Eurasian" is also somewhat lower, with the remaining replicates almost all placing the first ancestral population along the Pima branch instead. However, this is perhaps not surprising given evidence of European-related admixture in Pima; overall, our conclusions are unchanged, and the results appear quite robust to perturbations in the scaffold. See Figure 3.2A and the caption of Table 3.2 for descriptions of the parameters.

Table C.2. Mixture parameters for other populations modeled as two-way admixtures inferred with an alternative scaffold tree.

AdmixedPop	Branch1 + Branch2	# rep	α	Branch1Loc	Branch2Loc	MixedDrift
Daur	Anc. N-Eur. + She	264	0.225-0.459	0.005-0.052 / 0.151	0.002-0.014 / 0.016	0.014-0.024
	Anc. N-Eur. + Miao	213	0.235-0.422	0.005-0.049 / 0.151	0.002-0.008 / 0.008	0.014-0.024
Hezhen	Anc. N-Eur. + She	257	0.230-0.442	0.005-0.050 / 0.151	0.002-0.010 / 0.016	0.012-0.034
	Anc. N-Eur. + Miao	217	0.214-0.444	0.005-0.047 / 0.151	0.002-0.008 / 0.008	0.013-0.037
Oroqen	Anc. N-Eur. + She	336	0.284-0.498	0.010-0.052 / 0.151	0.003-0.015 / 0.016	0.017-0.036
	Anc. N-Eur. + Miao	149	0.271-0.476	0.007-0.046 / 0.151	0.002-0.008 / 0.008	0.018-0.039
Yakut	Anc. N-Eur. + Miao	246	0.648-0.864	0.004-0.018 / 0.151	0.005-0.008 / 0.008	0.032-0.043
	Anc. E-Asian. + Pima	71	0.917-0.973	0.008-0.020 / 0.045	0.022-0.083 / 0.083	0.028-0.042
	Anc. N-Eur. + She	161	0.664-0.865	0.004-0.018 / 0.151	0.003-0.017 / 0.017	0.030-0.043
Melanesian	Dai + Papuan	331	0.168-0.268	0.009-0.011 / 0.011	0.167-0.204 / 0.246	0.089-0.115
	Lahu + Papuan	78	0.174-0.266	0.005-0.034 / 0.034	0.167-0.203 / 0.244	0.089-0.118
Han	Karitiana + She	167	0.007-0.025	0.026-0.134 / 0.134	0.001-0.006 / 0.016	0.000-0.004
	She + Surui	54	0.971-0.994	0.001-0.006 / 0.016	0.017-0.180 / 0.180	0.000-0.003
	Anc. N-Eur. + She	65	0.021-0.080	0.004-0.105 / 0.152	0.001-0.007 / 0.016	0.000-0.003
	Pima + She	82	0.009-0.033	0.022-0.085 / 0.085	0.001-0.007 / 0.016	0.000-0.004

Mixture parameters inferred by *MixMapper* for non-European populations fit as two-way admixtures using an alternative unadmixed scaffold tree containing 11 populations: Yoruba, Mandenka, Mbuti Pygmy, Papuan, Dai, Lahu, Miao, She, Karitiana, Suruí, and Pima (see Figure C.2). The results for the first four populations are very similar to those obtained with the original scaffold tree, except that α is now estimated to be roughly 20% higher. Melanesian is fit essentially identically as before. Han, however, now appears nearly unadmixed, which we suspect is due to the lack of an appropriate northern East Asian population related to one ancestor (having removed Japanese). See Figure 3.2A and the caption of Table 3.2 for descriptions of the parameters; branch choices are shown that that occur for at least 50 of 500 bootstrap replicates. The “Anc. East Asian” branch is the common ancestral branch of the four East Asian populations in the unadmixed tree.

Table C.3. Mixture parameters for populations modeled as three-way admixtures inferred with an alternative scaffold tree.

Admixed2	Branch3[a]	# rep[b]	α_2[c]	Branch3Loc[d]	Drift1A[e]	Drift1B[e]	Drift2[e]
Druze	Mandenka	309	0.958-0.984	0.004-0.009 / 0.009	0.088-0.102	0.021-0.029	0.005-0.013
Palestinian	Mandenka	249	0.907-0.935	0.008-0.009 / 0.009	0.087-0.100	0.022-0.030	0.001-0.008
	Anc. W. Eurasian	92	0.822-0.893	0.050-0.122 / 0.246	0.102-0.126	0.000-0.019	0.011-0.023
Bedouin	Mandenka	303	0.852-0.918	0.006-0.009 / 0.009	0.086-0.101	0.022-0.030	0.007-0.019
Mozabite	Mandenka	339	0.684-0.778	0.006-0.009 / 0.009	0.095-0.112	0.010-0.021	0.018-0.032
	Yoruba	50	0.673-0.778	0.005-0.010 / 0.010	0.093-0.111	0.010-0.020	0.018-0.031
Hazara	Anc. East Asian	390	0.350-0.464	0.009-0.023 / 0.045	0.084-0.119	0.001-0.033	0.004-0.012
Uygur	Anc. East Asian	390	0.312-0.432	0.007-0.022 / 0.045	0.091-0.124	0.000-0.027	0.000-0.009

Mixture parameters inferred by *MixMapper* for populations fit as three-way admixtures using an alternative unadmixed scaffold tree containing 11 populations: Yoruba, Mandenka, Mbuti Pygmy, Papuan, Dai, Lahu, Miao, She, Karitiana, Suruí, and Pima (see Figure C.2). In all cases one parent population splits from the (admixed) Sardinian branch and the other from Branch3. All the parameters are quite similar to those obtained with the original scaffold with only some relative changes in bootstrap support among alternative topologies. See Figure 3.2B and the caption of Table 3.2 for further descriptions of the parameters; branch choices are shown that that occur for at least 50 of the 390 bootstrap replicates having the majority branch choices for the two-way Sardinian fit. The "Anc. East Asian" branch is the common ancestral branch of the four East Asian populations in the unadmixed tree.

Bibliography

Albrechtsen A, Nielsen F, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* 27:2534–2547.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410.

Bramanti B, Thomas M, Haak W, et al. 2009. Genetic discontinuity between local hunter-gatherers and Central Europe's first farmers. *Science* 326:137–140.

Brandon MC, Wallace DC, Baldi P. 2009. Data structures and compression algorithms for genomic sequence data. *Bioinformatics* 25:1731–1738.

Cavalli-Sforza L, Edwards A. 1967. Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics* 19:233–257.

Chakraborty R, Weiss K. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences* 85:9119–9123.

Chen G, Marjoram P, Wall J. 2009. Fast and flexible simulation of DNA sequence data. *Genome Research* 19:136–142.

Chen X, Li M, Ma B, Tromp J. 2002. DNACompress: fast and effective DNA sequence compression. *Bioinformatics* 18:1696–1698.

Chikhi L, Bruford M, Beaumont M. 2001. Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158:1347–1362.

Christley S, Lu Y, Li C, Xie X. 2009. Human genomes as email attachments. *Bioinformatics* 25:274–275.

Clark A, Hubisz M, Bustamante C, Williamson S, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15:1496–1502.

Davies R. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64:247–254.

Deorowicz S, Grabowski S. 2011. Robust relative compression of genomes with random access. *Bioinformatics* 27:2979–2986.

Der Sarkissian C, Balanovsky O, Brandt G, et al. 2013. Ancient DNA reveals prehistoric gene-flow from Siberia in the complex human population history of North East Europe. *PLoS Genet.* 9:e1003296.

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. 2004. Estimating the impact of prehistoric admixture on the genome of Europeans. *Molecular Biology and Evolution* 21:1361–1372.

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7:1–26.

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1:54–75.

Fletcher W, Yang Z. 2009. Indelible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution* 26:1879–1888.

Fujita P, Rhead B, Zweig A, et al. 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Research* 39:D876–D882.

Gravel S. 2012. Population genetics models of local ancestry. *Genetics* 191:607–619.

Gravel S, Henn B, Gutenkunst R, et al. 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* 108:11983–11988.

Green R, Krause J, Briggs A, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.

Gronau I, Hubisz M, Gulko B, Danko C, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics* 43:1031–1034.

Gross M. 2011. Riding the wave of biological data. *Curr Biol* 21:R204–R206.

Grumbach S, Tahi F. 1994. A new challenge for compression algorithms: Genetic sequences. *J. Inf. Process. Manage.* 30:875–886.

Haak W, Balanovsky O, Sanchez J, et al. 2010. Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. *PLoS biology* 8:e1000536.

Hammer M, Horai S. 1995. Y chromosomal DNA variation and the peopling of Japan. *American Journal of Human Genetics* 56:951–962.

Hammer M, Karafet T, Park H, Omoto K, Harihara S, Stoneking M, Horai S. 2006. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *Journal of Human Genetics* 51:47–58.

Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. 2011. Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome Research* 21:734–740.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. *Science* 326:1541–1545.

Huson D, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.

Huttenhower C, Hofmann O. 2010. A quick guide to large-scale genomic data mining. *PLoS Comput Biol* 6:e1000779.

Jarvis J, Scheinfeldt L, Soi S, et al. 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in western African Pygmies. *PLoS Genetics* 8:e1002641.

Kahn SD. 2011. On the future of genomic data. *Science* 331:728–729.

Keinan A, Mullikin J, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics* 39:1251–1255.

Keller A, Graefen A, Ball M, et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications* 3:698.

Kent WJ. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12:656–664.

Kircher M, Kelso J. 2010. High-throughput DNA sequencing - concepts and limitations. *BioEssays* 32:524–536.

Kozanitis C, Saunders C, Kruglyak S, Bafna V, Varghese G. 2011. Compressing genomic sequence fragments using SlimGene. *Journal of Computational Biology* 18:401–413.

Lander ES, Linton LM, Birren B, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology* 10:R25.

Laval G, Patin E, Barreiro L, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE* 5:e10284.

Lawson D, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genetics* 8:e1002453.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Ruan J, Durbin R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18:1851–1858.

Li J, Absher D, Tang H, et al. 2008b. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Lipson M, Loh P, Levin A, Reich D, Patterson N, Berger B. 2012. Efficient moment-based inference of admixture parameters and sources of gene flow. *In revision; arXiv preprint arXiv:1212.2555* .

Liti G, Carter DM, Moses AM, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* 458:337–341.

Loh P, Baym M, Berger B. 2012. Compressive genomics. *Nature Biotechnology* 30:627–630.

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233–1254.

Mäkinen V, Navarro G, Sirén J, Välimäki N. 2010. Storage and retrieval of highly repetitive sequence collections. *Journal of Computational Biology* 17:281–308.

Moorjani P, Patterson N, Hirschhorn J, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price A, Reich D. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics* 7:e1001373.

Moorjani P, Patterson N, Loh PR, et al. 2013. Reconstructing Roma history from genome-wide data. *PLoS ONE* 8:e58633.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press.

Nielsen R, Hubisz M, Clark A. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168:2373–2382.

Ohta T, Kimura M. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 68:571–580.

Pasaniuc B, Zaitlen N, Lettre G, et al. 2011. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genetics* 7:e1001371.

Patin E, Laval G, Barreiro L, et al. 2009. Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics* 5:e1000448.

Patterson N, Hattangadi N, Lane B, et al. 2004. Methods for high-density admixture mapping of disease genes. *American Journal of Human Genetics* 74:979–1000.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192:1065–1093.

Patterson N, Price A, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genetics* 2:e190.

Pickrell J, Patterson N, Barbieri C, et al. 2012. The genetic prehistory of southern Africa. *Nature Communications* 3.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* 8:e1002967.

Pinhasi R, Thomas M, Hofreiter M, Currat M, Burger J. 2012. The genetic history of Europeans. *Trends in Genetics* 28:496–505.

Pool J, Hellmann I, Jensen J, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Research* 20:291–300.

Pool J, Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* 181:711–719.

Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38:904–909.

Price A, Tandon A, Patterson N, Barnes K, Rafaels N, Ruczinski I, Beaty T, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* 5:e1000519.

Price A, Weale M, Patterson N, et al. 2008. Long-range LD can confound genome scans in admixed populations. *American Journal of Human Genetics* 83:132–135.

Pritchard J, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biology* 12:R19.

Quintana-Murci L, Quach H, Harmant C, et al. 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter–gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences* 105:1596.

Rasteiro R, Chikhi L. 2009. Revisiting the peopling of Japan: an admixture perspective. *Journal of Human Genetics* 54:349–354.

Reich D, Cargill M, Bolk S, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199–204.

Reich D, Green R, Kircher M, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–1060.

Reich D, Patterson N, Kircher M, et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics* 89:516–528.

Reich D, Thangaraj K, Patterson N, Price A, Singh L. 2009. Reconstructing Indian population history. *Nature* 461:489–494.

Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldman M. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating local ancestry in admixed populations. *American Journal of Human Genetics* 82:290–303.

Schatz M, Langmead B, Salzberg S. 2010. Cloud computing and the DNA data race. *Nat Biotech* 28:691–693.

Semino O, Passarino G, Oefner P, et al. 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* 290:1155–1159.

Sirén J, Marttinen P, Corander J. 2011. Reconstructing population histories from single nucleotide polymorphism data. *Molecular Biology and Evolution* 28:673–683.

Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert M, Götherström A, Jakobsson M. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* 336:466–469.

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt H, Torroni A, Richards M. 2010. The archaeogenetics of Europe. *Current Biology* 20:R174–R183.

Sousa V, Fritz M, Beaumont M, Chikhi L. 2009. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* 181:1507–1519.

Stratton M. 2008. Genome resequencing and genetic variation. *Nat Biotech* 26:65–66. 10.1038/nbt0108-65.

Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *American Journal of Human Genetics* 79:1–12.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:1.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

The International HapMap Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.

Tian C, Gregersen P, Seldin M. 2008. Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics* 17:R143–R150.

Tweedie S, Ashburner M, Falls K, et al. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucl. Acids Res.* 37:D555–559.

Venter JC, Adams MD, Myers EW, et al. 2001. The sequence of the human genome. *Science* 291:1304.

Verdu P, Austerlitz F, Estoup A, et al. 2009. Origins and genetic diversity of Pygmy hunter-gatherers from western Central Africa. *Current Biology* 19:312–318.

Wall J, Lohmueller K, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution* 26:1823–1827.

Wang J. 2003. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 164:747–765.

Xu S, Pugach I, Stoneking M, Kayser M, Jin L, et al. 2012. Genetic dating indicates that the Asian–Papuan admixture through eastern Indonesia corresponds to the Austronesian expansion. *Proceedings of the National Academy of Sciences* 109:4574–4579.

Yu Y, Degnan J, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* 8:e1002660.