Relatedness, admixture and the genetic history of Greenland-European contact

PhD thesis by Ryan Kele Waples

Submitted April 2019

Main supervisor: Ida Moltke
Co-supervisor: Anders Albrechtsen

Affiliations:
Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark

to dad

Preface

This thesis represents three years of work, from April 2016 to April 2019, at The Bioinformatics Centre, Department of Biology, University of Copenhagen, in Copenhagen, Denmark under the main supervision of Ida Moltke and co-supervised by Anders Albrechtsen.

A great aspect to a PhD in Denmark is that it requires a change of academic environment. For my exchange of environment, I spent three months in the lab of Garrett Hellenthal at University College London, London, UK.

Ryan Waples
April, 2019

Abstract

This thesis covers work on three basic concepts in population genetics: relatedness, admixture, and linkage disequilibrium, and contains both new statistical methods as well as application of recently published methods to better understand the recent human past in Greenland.

The first manuscript included in this thesis addresses a gap in the availability of methods to infer relatedness with limited data. Relatedness matters for all aspects of population genetics, but most methods to infer relatedness rely on the availability of population allele frequencies and accurate genotype data. We present a method that can infer close familial relationships without relying on estimates of population allele frequencies, and directly from low-depth (4x) sequencing data without genotype calling. It requires genetic data from only two individuals and has the potential to expand the number of studies able to infer relatedness despite limited data.

In the second manuscript we examine the history of contact between the Greenlandic Inuit and Europeans from a genetic perspective. The current population of Greenland has experienced substantial gene flow from Europe, but the European source countries of this ancestry was genetically unknown. There is a tight historical relationship between Greenland and Denmark, but there is also a history of Dano–Norwegian and German missionaries, Dutch whalers, as well as other European contact with Greenland. Using dense SNP array data from Greenlanders and Europeans we identify Denmark as the primary source of European ancestry in Greenland, and quantify the ancestry contribution from 14 different European countries. We discuss in detail how these results reflect the history of Greenland/European contact.

In the third and last manuscript we present a software tool for estimating linkage disequilibrium (LD) in admixed populations. The LD in a population is affected by many aspects of the population's history, including effective population size and past admixture. LDadmix estimates the two-locus haplotype frequencies within the source ancestries of a recently admixed population. These two-locus haplotype frequencies reflect the LD within each ancestry source prior to the admixture event. Through simulations and application to real data, we show that LDadmix can recover LD patterns in different admixture scenarios and also infer an elevated LD decay curve for the ancestral American ancestry, a signal that was previously masked by recent African and European admixture.

Together, these manuscripts highlight the continuing need for genetic methods that can be applied in the challenging and data-limited scenarios that will continue to be frequent in biology despite the recent expansion of available genetic data. They also highlight how new insights can be gained about different populations, like the Greenlandic, when such methods are developed and applied.

Dansk Resumé

Denne afhandling indeholder studier af tre basale populationsgenetiske koncepter: relatedness (et mål for hvor tæt folk er i familie), admixture (et fænomen hvor individer fra forskellige befolkninger får børn sammen) og LD (et mål for statisk afhængighed mellem genetiske loci). Afhandlingen præsenterer både nye statistiske metoder samt anvendelser af andre relativt nye metoder til at øge vores forståelse af den grønlandske befolknings historie.

Det første manuskript i denne afhandling adresserer det faktum, at der mangler metoder til at estimere relatedness baseret på begrænset genetisk data. Relatedness spiller en vigtig rolle i alle dele af populationsgenetik, men de fleste metoder til at estimere relatedness kræver adgang til populations-specifikke allelefrekvenser og genotype data af høj kvalitet. Vi præsenterer en ny metode, der kan bruges til at inferere tætte familierelationer uden at bruge populations-specifikke allelefrekvenser og som kan anvendes direkte på lav-dybde sekventeringsdata (4x) uden at kalde genotyper. Metoden kræver udelukkende genetisk data fra to individer og kan dermed potentielt udvide antallet af studier, hvor det er muligt at inferere relatedness.

I det andet manuskript undersøger vi den historiske kontakt mellem grønlandske inuit og europæere fra et genetisk perspektiv. I den nuværende grønlandske befolkning har mange ikke blot inuit forfædre men også europæiske forfædre, men det er ikke klart hvilke europæiske lande deres europæiske forfædre kommer fra. Der er et tæt historisk forhold mellem Grønland og Danmark, men historisk set har også nordmænd, tyske missionærer, hollandske hvalfangere og europæere fra andre lande haft kontakt med Grønland. Ved at analysere SNP array data fra grønlændere og europæere identificerede vi Danmark som det primære oprindelsesland for grønlændernes europæiske forfædre, og kvantificerer i hvor høj grad 14 forskellige europæiske lande har været oprindelseslande. Vi diskuterer derefter i detaljer, hvordan vores resultater passer med den historiske viden der er om grønlandsk-europæisk kontakt.

I det tredje og sidste manuskript, præsenterer vi et computerprogram til at estimere LD i en admixed befolkning. LD i sådanne befolkninger er påvirket af mange aspekter af befolkningens historie, inklusiv effektiv populationsstørrelse og admixture. LDadmix estimerer to-locus haplotypefrekvenser i hver af de ancestrale befolkninger til en nyligt admixed befolkning. Disse to-locus haplotypefrekvenser bliver derefter brugt til at estimere LD i disse ancestrale befolkninger inden admixture. Gennem simuleringer og analyser af rigtig data viser vi, at LDadmix kan genskabe ancestrale LD mønstre i forskellige scenarier og vi viser også, at der er højere LD i den ancestrale amerikanske befolkning, et signal der tidligere var maskeret af nylig admixture med afrikanere og europæere.

Samlet viser de tre manuskripter tydeligt at der – selv for basale koncepter som relatedness, admixture og LD - stadig er brug for nye genetiske metoder, der kan bruges i situationer hvor data er kompliceret eller begrænset. Det er situationer der vil blive ved med at opstå til trods for at der efterhånden er store mængder genetisk data tilgængelig. Manuskripterne viser også tydeligt hvordan sådanne nye metoder kan generere ny viden om forskellige befolkninger som for eksempel den grønlandske.

Reading guide

The primary content of this thesis is three manuscripts of which I am first author. One manuscript is already published, the other two manuscript are in preparation, and are presented here as drafts.

This thesis also contains a short introduction to each manuscript, located together in chapter 1.

These introductions each have five parts. I start with a brief background of the most important topics present in the manuscript. Next, I address the motivations behind the project and the goals of the manuscript, followed by a short summary of the results it presents. After that, I clearly state my contributions to the manuscript. I finish with a reflection on aspects of the study not fully covered in the manuscript and also consider further directions for the research.

After this three-part introduction, I have a few concluding remarks.

As an appendix, I have attached two additional papers I have co-authored during my PhD. These represent some of the further work I have done during my PhD, but on projects where I played a supportive, rather than lead role.

List of publications

Ryan K. Waples, Anders Albrechtsen, and Ida Moltke. 2019. "Allele Frequency-Free Inference of Close Familial Relationships from Genotypes or Low-Depth Sequencing Data." Molecular Ecology 28 (1): 35–48.

Ryan K. Waples, Aviaja Lyberth Hauptmann, Inge Seiding, ..., Garrett Hellenthal, Torben Hansen, Anders Albrechtsen, Ida Moltke. "Where did the European ancestors of the Greenlanders come from?" (in preparation)

Ryan K. Waples, Anders Albrechtsen, and Ida Moltke. "Estimating linkage disequilibrium in admixed populations" (in preparation)

Other publications since start of PhD (April 2016)

In appendix

Arthur Gilly, Daniel Suveges, Karoline Kuchenbaecker, Martin Pollard, Lorraine Southam, Konstantinos Hatzikotoulas, Aliki-Eleni Farmaki, Thea Bjornland, **Ryan K. Waples**, Emil V. R. Appel, Elisabetta Casalone, Giorgio Melloni, Britt Kilian, Nigel W. Rayner, Ioanna Ntalla, Kousik Kundu, Klaudia Walter, John Danesh, Adam Butterworth, Inês Barroso, Emmanouil Tsafantakis, George Dedoussis, Ida Moltke, and Eleftheria Zeggini. "Cohort-wide Deep Whole Genome Sequencing and the Allelic Architecture of Complex Traits." Nature Communications 9, no. 1 (2018).

David W. G. Stanton, Peter Frandsen, **Ryan K. Waples**, Rasmus Heller, Isa-Rita M. Russo, Pablo A. Orozco-Terwengel, Casper-Emil Tingskov Pedersen, Hans R. Siegismund, and Michael W. Bruford. "More Grist for the Mill? Species Delimitation in the Genomic Era and Its Implications for Conservation." Conservation Genetics 20, no. 1 (2019): 101-13.

Masters Thesis

Ryan K. Waples, James E. Seeb, and Lisa W. Seeb. "Congruent Population Structure across Paralogous and Nonparalogous Loci in Salish Sea Chum Salmon (Oncorhynchus Keta)." Molecular Ecology 26, no. 16 (2017): 4131-144

Other

Ryan K. Waples, Wes. A. Larson, and Robin S. Waples. "Estimating Contemporary Effective Population Size in Non-model Species Using Linkage Disequilibrium across Thousands of Loci." Heredity 117, no. 4 (2016): 233-40

Garrett J. Mckinney, **Ryan K. Waples**, Carita E. Pascal, Lisa W. Seeb, and James E. Seeb. "Resolving Allele Dosage in Duplicated Loci Using Genotyping-by-sequencing Data: A Path Forward for Population Genetic Analysis." Molecular Ecology Resources 18, no. 3 (2018): 570-79.

Garrett J. Mckinney, **Ryan K. Waples**, Lisa W. Seeb, and James E. Seeb. "Paralogs Are Revealed by Proportion of Heterozygotes and Deviations in Read Ratios in Genotyping-by-sequencing Data from Natural Populations." Molecular Ecology Resources 17, no. 4 (2016): 656-69.

Michael W. Ackerman, Brian K. Hand, **Ryan K. Waples**, Gordon Luikart, Robin S. Waples, Craig A. Steele, Brittany A. Garner, Jesse Mccane, and Matthew R. Campbell. "Effective Number of Breeders from Sibship Reconstruction: Empirical Evaluations Using Hatchery Steelhead." Evolutionary Applications 10, no. 2 (2016):

Contents

1

Introduction

1.1 Allele Frequency-Free Inference of Relationships

1.1.1 Background

Relatedness

Relatedness matters for all aspects of population genetics, from applied and practical to theoretical and statistical. In conservation genetics, managed breeding programs that are designed to avoid inbreeding must be aware of relatedness (e.g. Putnam and Ivy, 2014). In studies of ancient human populations, understanding the relatedness of individuals buried together can shed light on social organization and migration patterns (e.g. Amorim et al., 2018). In all species, the relatedness among a sample of individuals can inform about the relative benefit of further sequencing or sampling efforts, and depending on the study design, related individuals may be desired, as in sibling-based heritability studies (e.g. Athanasiadis et al., 2019) or unwelcome, when estimating population structure (Pritchard et al., 2000). Many statistical methods in population genetics will return spurious results if they fail to account for relatedness (Voight and Pritchard, 2005). And finally, the recent popularity of at-home genetic ancestry companies is at least in part due to the desire of individuals to find relatives.

There are multiple definitions of relatedness that can be useful depending on the context. A pedigree (Figure 1.1A) specifies a degree of (pedigree) relatedness for all the individuals it includes. Pedigree relatedness addresses the familial relationship categories we are most familiar with: grandparent, aunt-uncle, second cousin, as well as more distant ones such as 8th cousins, once-removed. Recently, population genetics has often found more use for genomic measures of relatedness which use genetic similarity to quantify relatedness because they better reflect important biological and evolutionary processes (Speed and Balding, 2015; Kardos et al., 2015; Wang, 2016). Genomic measures of relatedness can be useful in a number of ways. They can can be used directly, to look at inbreeding, or as in genetic association tests to correct for correlation in phenotypes due to shared ancestry. They can also be used to infer pedigree relationships between individuals, as we will discuss further below.

There are multiple possible measures of genetic similarity available to estimate genomic relatedness, including correlation of genotypic values (e.g. Yang et al., 2010), kinship/coancestry coefficients (Wright, 1922), identity-by-descent tracts (e.g. Browning and Browning, 2013), as well as coalescent estimates of time to most recent common ancestor (e.g. Speidel et al., 2019). Many of these rely on the concept of identity-by-descent (IBD) (Thompson, 2013). IBD is the concept of *recent* shared ancestry, and is a useful concept in part because it allows biologists to quantify relatedness in a flexible manner. Inherent in the name, identity-by-descent, is the idea that the shared identity is descended from somewhere, a chosen reference population or ancestor. The interpretation of IBD changes based on the choice of how this reference is chosen so that the concept of IBD can be applied in a variety of ways (e.g. Staples et al., 2014; Palamara et al., 2012; Browning and Thompson, 2012; Albrechtsen et al., 2010).

Below, and in manuscript 1, we utilize the k-coefficients of Cotterman (1940) to describe relatedness between pairs of individuals. They specify the degree of genomic relatedness between two non-inbred diploid individuals with 3 coefficients that sum to 1, where k_0, k_1, k_2 is the probability that the pair of individuals have 0, 1, or 2 alleles IBD at a random site on the genome, respectively. Pairs of diploid individuals with different pedigree relationships have different expected values of these k-coefficients

(see Table 1.1), and estimates of these coefficients, are often used to infer pedigree relatedness. The k-coefficients can also be collapsed into a single estimate of kinship or coancestry ($\theta = k_1/4, k_2/2$) (Lynch and Walsh, 1998) which can be used for pedigree relationship inference in much the same manner (e.g. Manichaikul et al., 2010).

Table 1.1: Expected $K = (k_0, k_1, k_2)$ for different relationship categories. Reproduced from the supplement of manuscript 1.

Relationship	k_0	k_1	k_2
Monozygotic twins (MZ)	0	0	1
Parent-offspring (PO)	0	1	0
Full siblings (FS)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Half siblings/avuncular/grandparent-grandchild (HS)	$\frac{1}{2}$	$\frac{1}{2}$	0
First cousins (C1)	$\frac{3}{4}$	$\frac{1}{4}$	0
Second cousins (C2)	$\frac{15}{16}$	$\frac{1}{16}$	0
Unrelated (UR)	1	0	0

The connection between pedigree and genomic relatedness is complicated by randomness inherent in the process of recombination. The inheritance processes that generates IBD leads to IBD occurring in blocks along the genome, which are broken by recombination events. But, due to the stochastic process of recombination, the realised amount of IBD that is present between pairs of individuals with the exact same pedigree relationship can vary by quite a lot. In practice this means that there are fundamental limits to the ability to reconstruct distant pedigree relationships (Hill and Weir, 2011). For close relationships this approach works well with sufficient data (e.g. Manichaikul et al., 2010). These concepts are illustrated in Figure 1.1. Figure 1.1A, shows an extended pedigree, with two different pedigree relationships highlighted with boxes, first cousins and avuncular. In Figure 1.1B,D, there are results of simulations of the IBD process in the two pairs of related individuals with relationships indicated by the pedigree, with regions of the genome colored by the IBD status shared by the pair of individuals. Figure 1.1C, shows how the fraction of the genome with IBD 1 (i.e. k_1) varies across 500 replicate simulations of the inheritance process for each relationship. The simulated genome is an approximation to the human genome, with 22 chromosomes with lengths taken from a human genetic map. Each generation the number of recombination events were selected from a Poisson distribution, enforcing at least one event per chromosome.

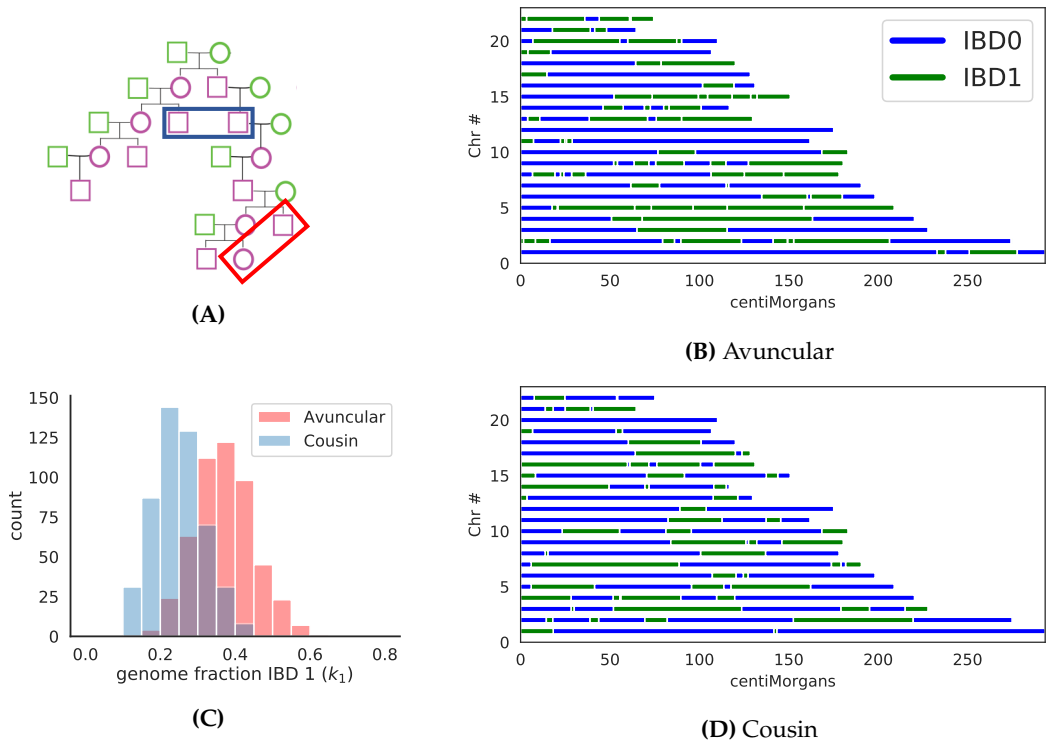


Figure 1.1: **A**) Example pedigree with two highlighted relationships: first cousins (blue box) and avuncular (e.g. uncle-nephew) (dark blue box). **B**) Realization of the IBD process for an avuncular relationship for a simulated human genome. **C**) Histogram of the fraction of the genome that is IBD 1 (k_1) across 500 replicate simulations for each of the example relationships. **D**) Realization of the IBD process for a first cousin relationship for a simulated human genome. The pedigree figure in (A) was adapted from Li et al. (2014)

Estimating relatedness

Unfortunately, the IBD status between a pair of individuals, and thus also the k-coefficients, cannot be observed directly and must be inferred from genetic data. What can be observed in genetic data is identity-by-state (IBS). Two or more genetic sequences are IBS in sites where they have the same allele. IBS is informative about IBD because a pair genetic sequences that are IBD are much more likely to be IBS than a random pair sequences from two unrelated individuals from the same population. The basic idea behind IBD inference for a pair of individuals is therefore to look for genetic sequence identity above what is expected for a pair of unrelated individuals. This undertaking is slightly complicated by the fact that, despite the name, identity-by-descent does not completely indicate identity-by-state; mutations since the choice of the reference population can allow IBD without IBS (although there are interpretations of IBD that do not accommodate this view). Other confounding factors, such as genotyping error, must also be addressed. Fortunately, these effects are relatively minor in most cases, especially for SNPs due to their low mutation rate (Scally, 2016).

There are two main concepts involved in using observed IBS patterns to infer IBD status. First, to identify elevated levels of IBS, the expected amount of IBS for an unrelated pair of individuals must be known. The expected IBS for a pair of unrelated diploid individuals can be based on the allele

frequencies in the population, here shown for a single locus:

$$I = 2q^4 + 2p^4 + 8p^2q^2 + 4pq^3 + 4p^3q. \tag{1.1}$$

Where I is the expected number of alleles the two individuals share at a locus with allele frequencies p and q=(1-p). Genome-wide IBS in excess of expectations can be modeled as occurring due to recent shared ancestry (i.e. IBD). Maximum-likelihood estimates of the previously introduced k-coefficients can be achieved from genotype data (G) and population allele frequencies with the likelihood of the k-coefficients equal to the probability of the genotype data (G) given the k-coefficients $K = [k_0, k_1, k_2]$ (Thompson, 1975; Choi et al., 2009):

$$L(K) = PR(G|K) = \sum_j PR(G|n_{ibd} = j)k_j \tag{1.2}$$

where $n_{ibd} \in [0, 1, 2]$ is the number of alleles that are IBD at each site and $PR(G|n_{ibd})$, the probability of the genotype data conditional on the IBD status depends on the allele frequencies in the population (for details see the supplement to manuscript 1, or see Table 2 of Purcell et al. (2007)). This and similar methods clearly rely on accurate estimates of population allele frequencies.

As discussed briefly above, the inheritance processes that generates IBD occur along the genome, so that adjacent sites have a correlated IBD status. This means that we do not expect sites that are IBD to be randomly scattered across the genome, instead IBD sites should appear in clusters, often called IBD tracts, with regions of IBD interspersed with non-IBD regions. In the absence of IBD, IBS is unlikely to occur by chance in long consecutive regions. So if the order of sites is known, this idea can be leveraged to infer IBD without utilizing population allele frequencies by looking for extended regions of IBS (e.g. Gusev et al., 2009; Stevens et al., 2011). However, as these methods rely on allelic identity without conditioning on allele frequencies, they often use arbitrary thresholds that define how clusters of matching alleles are interpreted. Of course, these two concepts used to characterize IBD can also be combined, as is done in some IBD inference methods (e.g. Browning and Browning, 2013).

There are a huge number of methods and software programs available to estimate genomic relatedness based on the above ideas. They have a wide variety of data requirements, utilize a range of statistical methods, return different measures of relatedness, and have different goals. I will not go through all of these in detail here, but I have collected a representative sample of software for estimating relatedness from SNP genotype data (Table 1.2) and for methods estimating relatedness from next-generation sequence data (Table 1.3). These tables aim to provide summaries of the input requirements (i.e. are allele frequencies used or the order of sites) and a sense of how the methods address relatedness.

Table 1.2: Selected SNP genotype-based methods for relatedness inference

Method		Utilizes		Infers		
		allele frequencies	site order	kinship	pedigree relationship	IBD tracts
(Chang et al. 2015)	PLINK	yes	no	yes	no	no
(Albrechtsen et al. 2009)	Relate	yes	yes	yes	no	yes
(Manichaikul et al. 2010)	KING-homo	yes	no	yes	yes	no
(Yang et al. 2011)	GCTA GRM	yes	no	no	no	no
(Conomos et al. 2016)	PC-relate	yes[1]	no	yes	no	no
(Stevens et al. 2011)	kcoeff	no	yes	yes	no	no
(Gusev et al. 2009)	GERMLINE	no	yes	no	no	yes
(Li et al. 2014)	GRAB	no	yes	no	yes	no
(Manichaikul et al. 2010)	KING-robust	no	no	yes	yes	no
(Lee 2003)	Lee	no	no	no	no	no
(Browning and Browning 2013)	Refined IBD	yes	yes	yes[2]	no	yes

Table 1.3: Sequencing-based methods for relatedness inference

Method		Utilizes		Infers	
		allele frequencies	site order	kinship	pedigree relationship
(Dou et al. 2017)	SEEKIN	yes	(imputation)	yes	no
(Korneliussen and Moltke 2015)	NGSrelate	yes	no	yes	no
(Kuhn et al. 2018)	READ	no[3]	yes	no	yes
(Theunert et al. 2017)	relcoas	yes	no	no	no

There are a few things to take away from such a survey. One, there are quite a number of methods to infer relatedness with SNP genotype data, as Table 1.2 could be much longer. But, more importantly, there are fewer good options for use with sequencing data, and especially there are very few options that can be applied without allele frequencies (Table 1.3) for sequencing studies with limited sample size. I will elaborate further on this point below.

Finally, there are two important aspects to inferring relatedness that I have glossed over, but that warrant a brief mention. First, diploid individuals have two copies of each chromosome, and if the two copies of a chromosome within a single individual are IBD, we refer to this as inbreeding. This occurs due to IBD that is present between the parents of the considered individual. Depending on the species, population, and pedigree, inbreeding can be important to account for during the inference of relatedness. It is

[1] individual allele frequencies
[2] with provided script
[3] requires an another unrelated individual

possible to extend the k-coefficients of Cotterman (1940) to a system that allows for inbreeding (Jacquard, 1972), I will briefly discuss this again below.

Second, all of my descriptions above assume the pair of individuals originates from the same homogenous population. However, natural populations commonly experience immigration and gene flow, so that the assumption of homogeneity is often not reasonable. If the two individuals originate from different populations, or have different ancestry compositions, allele-frequency based estimates of relatedness, including the k-coefficients and the kinship coefficient may be affected. This occurs because the allele frequencies in a single population no longer reflect the expected amount of IBS between a pair of individuals with no IBD. The result can be biased estimates of relatedness, and so it is important to be aware of and account for admixture. A common approach to infer relatedness in cases of admixture is to replace population allele frequencies with individual-specific allele frequencies (Pritchard et al., 2000; Hao et al., 2016), and then to estimate kinship accounting for the different allele frequencies in each individual (e.g. equation 4 of Conomos et al., 2016). This approach can work well, but requires the accurate estimation of individual allele frequencies, just as the homogenous population case relied on accurate estimates of the population allele frequencies.

However, while inbreeding and admixture can affect relatedness estimates and may be present in many datasets, practically they are often ignored. In fact, only a few of the methods listed in tables 1.2 and 1.3 can handle admixture or inbreeding, while the rest are not designed to account them.

1.1.2 Motivation and goals for manuscript 1

Given a set of unrelated individuals from a single population, allele frequencies are commonly estimated with simple maximum likelihood methods, and the accuracy of these estimates depends on sample size and a number of other factors I won't cover here. So for many reasons, accurate estimates of population allele frequencies may not be available. This can especially be a problem in studies of non-model organisms or ancient samples, as they often have very limited sample sizes. Without prior knowledge of allele frequencies, and with a limited sample size, estimates of relatedness that rely on allele frequencies perform much worse, (e.g. Wang, 2017; Theunert et al., 2017), and the number of applicable methods is substantially reduced (see Tables 1.2,1.3).

This is important because the impacted studies include both non-model organisms and ancient DNA. In these studies there are often a low number of samples available and the few samples that are available are only sequenced to low depth, making not only allele frequency estimation, but also genotype calling difficult.

In studies with limited sample size, but with dense SNP genotype data and a reference genome, methods that rely on the spatial pattern of IBS along the genome (e.g. GERMLINE, kcoeff, GRAB, Table 1.2) can be applied without allele frequencies. However, they have not been widely adopted for this use, likely due to the high data requirements and potential difficulty interpreting their output. Also, there is a bit of a mismatch between these methods and a lack of allele frequency information, as access to dense SNP genotype data and a reference genome are usually associated with large studies of current-day populations in well-studied species, where it is likely possible to achieve reasonable estimates of allele frequencies.

Motivated by a lack of appropriate methods, Monroy Kuhn et al. (2018) recently provided a method to estimate relatedness between a pair of ancient individuals, each represented by pseudo-haploid genotypes formed by sampling an allele from each site covered by a sequencing read. This method breaks the genome into 1Mb windows, and evaluates IBS within each window. In addition to the observed IBS pattern between the individuals, it also requires calibration of the degree to which IBS implies IBD. In frequency-based methods, this would be provided by the population allele frequencies. Rather than utilize allele frequencies directly, mean IBS between a set of unrelated individuals is used instead. This is an appealing approach, but it relies on prior knowledge of and access to comparable genetic data for a group of two or more unrelated individuals, as well as a contiguous reference genome to form the windows.

There is a clear a lack of methods for estimating relatedness with limited data and limited genetic resources. Motivated by this observation we sought to address this gap by developing a method that could be:

- Applied to genetic data from a pair of individuals, without external allele frequencies or access to additional individuals
- Applied to low or moderate depth sequencing data, as well as SNP genotype data
- Could be applied to incomplete reference genomes or contigs
- Robust to SNP ascertainment, and would work without prior knowledge of variable sites

1.1.3 Results

In manuscript 1, we presented a method that meets the above goals, drawing in part on previous work that also addressed relatedness in challenging scenarios. In some ways, the method could be considered an extension of ideas first presented in Lee (2003), a paper that presented an elegant binary test for related vs not-related based only on SNP data for a pair of individuals, without requiring estimates of allele frequencies. We also adopted a kinship estimator from Manichaikul et al. (2010), KING-robust, that was developed to be robust to latent population structure, and show how it is directly applicable to the issue at hand.

The basic logic of our method is shown in Figure 1.2, for a full explanation see Figure 1 in manuscript 1. Here we extend the Hardy-Weinberg expectation for genotypes in a single individual to a pair of individuals from the same population and construct a two-dimensional site-frequency spectrum (2d SFS). This 2d SFS presents one way of summarizing IBS between a pair of individuals. Figure 1.2A, shows the expected 2d SFS for a pair of unrelated individuals at a single di-allelic site, where allele 1 has frequency p, and allele 0 has frequency $q = 1 - p$. In Figure 1.2B, we show a diagram of how these values are expected to change if the individuals are related, i.e. have non-zero k_1 or k_2 coefficients. The expected values for each site in the genome can be combined to generate a genome-wide summary of IBS for the pair of individuals. This genome-wide IBS pattern is informative about IBD and thus in turn about pedigree relatedness. In the manuscript we formalized these basic observations using mathematical derivations assuming no admixture or inbreeding.

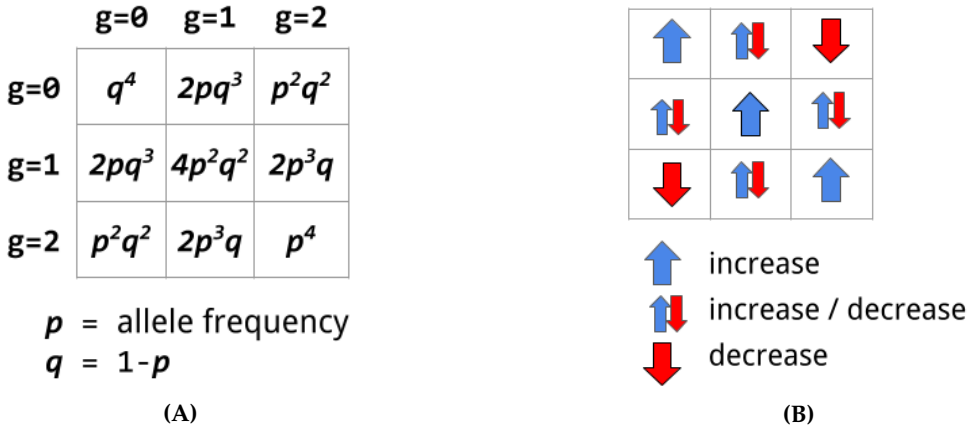


Figure 1.2: Conceptual introduction to the effect of relatedness on the 2d SFS for a pair of individuals from the same population. **A**) shows the expected 2d SFS for a pair of unrelated individuals at a single di-allelic site, with allele 1 having frequency p, and allele 0 having frequency $q = 1 - p$. The genotypes of the two individuals $g \in [0, 1, 2]$ are arrayed on the two axes. **B**) shows how these values are expected to change if the individuals are related, i.e. if they have non-zero k_1 or k_2 coefficients.

We then evaluated multiple ways to estimate this 2d SFS directly from sequencing data from a pair of individuals – and thus without calling genotypes – and showed how three different summary statistics calculated from this 2d SFS can be used to infer close familial relationships.

Using real data from the HGDP (Rosenberg, 2006) and the 1000 Genomes (1000 Genomes Project Consortium et al., 2015), we tested the method on both SNP array and whole-genome sequencing data, and demonstrated that it could recover previously reported pedigree relatives along a relatedness gradient from parent-offspring to first cousins. The depth of the whole-genome sequencing data was too low to confidently call genotypes, and we showed that genotype calling with this data severely impacted downstream estimates of relatedness, like the KING-robust kinship estimator. Furthermore, by downsampling the whole-genome sequencing data in a way that imitated restriction-site associated DNA approaches (RADseq) (Baird et al., 2008), we also showed how the method was robust to a reduced sequencing effort. Finally, using real and simulated data, the method was demonstrated to be robust to a wide array of ascertainment schemes as well as recent population size changes.

1.1.4 My contributions

For Manuscript 1, I performed all analyses and simulations, made all of the figures, and contributed to study design. An initial version of Figure 1 was made by Anders Albrechtsen. The manuscript writing was shared between me and Ida Moltke. Derivations in the supplement are by Ida Moltke.

1.1.5 Conclusions and future directions

The estimation of pairwise relatedness or genetic similarity is the first step of most genetic analyses, either because it is intrinsically of interest, or because it will aid in the interpretation of all downstream

results. For many studies in modern humans and other well-studied species these analyses have become even more important as sample size grow and the presence of relatedness becomes inevitable (Shchur and Nielsen, 2018). In contrast, manuscript 1 focuses on studies that are not drowning in data and are instead short on resources, either in terms of access to samples, sequencing coverage or budget, or access to genomic resources, and tries to allow a broader range of studies to conduct analyses of relatedness than was possible before.

Despite the positive results in manuscript 1, there is certainly room for further work in the area of inferring pairwise relatedness without external allele frequencies. Below, I will briefly discuss two broad areas that I feel are potential avenues for increasing the applicability and interpretability of the method. First, I will discuss the estimation of pairwise IBS from sequencing data, and then I will move on to discuss possible ways to better interpret IBS patterns for a pair of individuals of unknown relationship in the case of complicating factors such as inbreeding and admixture.

Correctly accounting for the statistical uncertainty of the genotype in low and moderate depth next-generation sequencing data is not an easy task. Some of the most powerful methods to deal with this uncertainty leverage data across many individuals (DePristo et al., 2011), but again manuscript 1 is focused on a more restricted case with only access to genetic data for a pair of individuals. We did not investigate in detail all of the potential pitfalls when estimating pairwise IBS for a novel pair of individuals, but we did report that genome-wide mappability filters ((e.g. Derrien et al., 2012), and the choice of genotype quality error model, as implemented in ANGSD (Korneliussen et al., 2014), both made a meaningful difference in our results. Investigating this in further detail was beyond the scope of the manuscript, but advances in the production, alignment, and assembly of sequencing data will all be relevant to improving these estimates.

The sequencing data we used to evaluate the method was approximately 4x depth. This was too low to call genotypes, and we showed that genotypes called from this data produced nonsense estimates of relatedness. It would certainly be worthwhile to investigate the sensitivity of the method to lower sequencing depths, which would make the estimation of pairwise IBS more difficult. A chief concern with lower sequencing depths is maintaining the ability to correctly recover the rate of shared heterozygous genotypes. Shared heterozygous sites are very informative about relatedness as they show both that the site is variable in the population and that the pair of individuals have the maximum local IBS value at this site. In contrast, a shared homozygous genotype also has the maximum local IBS value, but fails to provide any evidence that alleles at this site are segregating in the population. A minimum of four sequencing reads, two from each individual, is needed to clearly support the presence of a shared heterozygous genotype. A depth of four is not necessary at every site, but the presence of some sites with this depth are likely importantant for estimating relatedness with this method.

Application to ancient samples presents it own set of challenges. We did not test the method in manuscript 1 on ancient samples, but it is an area of future interest. It may be possible to adapt IBS estimation methods to account for some of the unique challenges of working with ancient samples, such as post-mortem damage to DNA (Handt et al., 1996). In the supplement to manuscript 1, we briefly noted the inclusion in ANGSD of methods that can estimate IBS for each of the 100 possible pairwise genotypes for the four bases of DNA (10 distinct genotypes in each individual). This has the potential to improve estimation of IBS for ancient DNA, as it can help account for the excess of C→T (G→A) changes due to DNA degradation. Often, ancient samples are treated as pseudo-haploid (e.g. Monroy Kuhn et al., 2018)

due to a low sequencing depth, but manuscript 1 offers a different approach that may be helpful in the future.

There are certainly also some important biological concerns that warrant further consideration. In manuscript 1 we specifically choose to only focus on pairs of individuals with no sign of inbreeding or admixture. This obviously excludes many studies outright, and of course it is certainly difficult to rule out inbreeding or admixture with limited data. We note that using an estimated pairwise 2d SFS it is possible to directly compare the heterozygosities of the two individuals and, with some assumptions, estimate the inbreeding coefficients. This is potentially very useful as inbreeding can be intrinsically interesting and can also bias estimates of relatedness if not accounted for.

Admixture also has the potential to bias estimates of relatedness, and has been shown to be problematic for the KING-robust kinship estimator (Manichaikul et al., 2010), due to the assumption of Hardy–Weinberg equilibrium. And since our method is in part based on the KING-robust kinship estimator this issue is also relevant to our method. We did not investigate the effects of admixture in manuscript 1, but the summary statistics we presented are sensitive to differential admixture between the individuals. It is possible to use the statistics in our method for quantifying the genetic differentiation between the ancestries of the two individuals, as noted by Manichaikul et al. (2010), but addressing pedigree relatedness at the same time is difficult. In the same vein, Lee (2003) suggested a test for different ancestries across a pair of individuals, essentially testing if the genotypes of the two individuals seemed be drawn from a single set of allele frequencies. The next step of actually extending the method to allow estimation of relatedness among admixed individuals without allele frequencies is left for a further study and may not be even possible without using the idea of IBD tracts. However, it should be noted that if the pair of individuals have the same admixture proportions, as may be common with older admixture events, the methods in manuscript 1 should apply directly.

To conclude, I will quickly note an interesting parallel of the method presented in manuscript 1 with the relatedness estimates reported for one of the largest studies to date, the UK biobank, which calculated relatedness between approximately 10^{11} pairs of individuals (Bycroft et al., 2018). Both papers utilized the KING-robust kinship estimator (Manichaikul et al., 2010), and noted it did not require specification of allele frequencies. It is interesting to see convergence of methods at such different scales.

1.2 Where did the European ancestors of the Greenlanders come from?

1.2.1 Background

A brief genetic history of the Greenlandic Inuit

Greenland is the largest island in the world and, due to its remote location and challenging arctic environment, it was first peopled circa 2500 BC. The present inhabitants of Greenland are the Greenlandic Inuit, descendants of the Thule culture that expanded out of Alaska and arrived in Greenland sometime before the 14th century CE (Gulløv, 2008; Friesen and Arnold, 2008). The Thule culture utilized innovative ways of living and thriving in the arctic, including new hunting and sailing technologies that allowed better use of the scarce resources in the challenging climate, especially marine mammals. The Inuit arrived in Northwest Greenland and expanded along the southern coast, eventually reaching East Greenland and Northeast Greenland (Sørensen and Gulløv, 2012; Gulløv, 2008). In 2019, the current population of Greenland is approximately 56000 (Statistics Greenland, accessed April 2019).

Previous inhabitants of Greenland include the Saqqaq and Dorset cultures (Early, Middle and Late) (Grønnow and Sørensen, 2006), as well as the Norse Vikings. There is evidence of interaction between the Late Dorset and the Thule Inuit in Greenland (Gulløv, 2008) and some genetic evidence admixture between the Inuit and people related to the Dorset, but it may have occurred in the Old World, prior to contact in Greenland (Raghavan et al., 2014).

Despite an apparent temporal overlap in Greenland prior to the end of the Norse period in Greenland in approx 1450 CE, the extent of interaction between the Inuit and the Vikings is still an area of study (Gulløv, 2008; Golding et al., 2011). There is evidence of some trade interactions between the Vikings and Native Greenlanders, but the Vikings and other Native Greenlanders remained distinctly economically independent (Gulløv, 2008). There is no clear genetic evidence of gene flow from the Vikings into the Inuit (Raghavan et al., 2014; Moltke et al., 2015).

Starting in the 16th century there was a new period of European contact with Greenland, as Europeans came to Greenland as explorers, whalers, missionaries, traders, and colonizers. In the 16th century, explorers from England searching for the Northwest passage were some of the first Europeans to arrive in Greenland after the Norse period ended in approximately 1450 CE (Gulløv, 2008; Frandsen et al., 2017). In the 17th and 18th centuries, whaling near Greenland was a big European economic activity, with fleets from the Netherlands, England, Denmark-Norway and other countries carrying sailors from many regions of Europe to Greenlandic harbors. In 1720's and 1730's, missionaries from Denmark-Norway and Germany arrived and established missions and religious traditions that continue to the present. In 1751, Denmark-Norway claimed a colonial trade monopoly with Greenland, with a goal towards excluding other nations access. Greenland remained a formal colony of Denmark until the 1950s, when Greenland and its population joined Denmark. In 1979 Greenland gained its own parliament and in 2009 it established self-government, though it remains part of the Kingdom of Denmark.

Previous work has characterized the genetic and demographic history of the Greenlandic Inuit. The genomes of the Greenlandic Inuit have signatures of adaptation to the arctic environment at loci in-

volved in fat metabolism (Fumagalli et al., 2015). Their genomes also show the effects of an "extreme and prolonged population bottleneck" (Pedersen et al., 2017), with high levels of LD (Moltke et al., 2015; Pereira et al., 2015) and a flattened site-frequency spectrum (Pedersen et al., 2017). Estimates of F_{ST} between the Greenlandic Inuit ancestry and other world populations are large, with estimates of 0.12 to the Han Chinese and 0.16 to Europe (Moltke et al., 2015). This work has both informed about migration patterns within Greenland (Moltke et al., 2015), and established the relationship of the Greenlandic Inuit to other world populations (Raghavan et al., 2014).

Previous work has also shown that while the current day population of Greenland draws most of its ancestry from the Inuit, it also has a significant amount of genetic ancestry from Europe, due to the history of contact and colonization (Bosch et al., 2003; Rasmussen et al., 2010; Pereira et al., 2015; Moltke et al., 2014). Moltke et al. (2015) estimated that the current population of Greenland derives approximately 75% percent of their ancestry from the Inuit and the remaining 25% from Europeans. The European ancestry in Greenland shows strong signals of a male sex bias (Bosch et al., 2003; Moltke et al., 2015), and also varies by location in Greenland, with the lowest rates of European ancestry occurring in North and East Greenland (Moltke et al., 2015).

Haplotype-based methods

An aspect of the genetic history of Greenland that has received less study are the sources of the European ancestry in the present day Greenland population. Europeans from many different countries have come to Greenland, but we don't have genetic evidence of where in Europe the European ancestry in Greenland is from. This is likely due to a lack of methods that have power to distinguish between the genetically very similar, potential European ancestry sources. Recently, a number of studies have had success resolving fine-scale genetic structure using haplotype-based methods to quantify genetic structure. Across the world, these methods have been applied to human populations in the British Isles (Leslie et al., 2015), Latin America (Chacón-Duque et al., 2018), Siberia (Sikora et al., 2018), Finland (Martin et al., 2018), The Democratic Republic of the Congo (van Dorp et al., 2019), and The Iberian Peninsula (Bycroft et al., 2018), and have revealed the genetic makeup of modern populations with remarkable resolution, as well as helped uncover and describe past patterns of human migration and admixture.

Unlike methods like STRUCTURE (Pritchard et al., 2000), that assume each locus provides independent information about population structure, haplotype-based methods utilize information present in the patterns of linkage disequilibrium and use haplotype-based measures of genetic similarity to identify shared ancestry. These methods have the ability to resolve subtle degrees of genetic structure that methods that assume independence between loci fail to capture (Lawson et al., 2012). In particular, this presents an opportunity to investigate the process of admixture in great detail (e.g. Hellenthal et al., 2014).

Lawson et al. (2012) presented a haplotype-based method, CHROMOPAINTER, that utilizes a hidden Markov model (HMM) that statistically reconstructs ("paints") the phased haplotypes of a target individual by reconstructing it as a mixture of reference haplotypes, following Li and Stephens (2003). Given a target haplotype and a set of reference haplotypes with the same sites, CHROMPAINTER tries to infer the best matching reference haplotype at each site in the target haplotype. Transition rates between adjacent genomic loci are given by a local recombination rate, and local allelic mismatches between the

target and reference haplotypes are accommodated with a mutation parameter. The path of reference haplotypes that is used to reconstruct the target is an approximation to the genealogical nearest neighbors along the target haplotype (Lawson et al., 2012). This reconstructed path, and more generally the expected frequency of each reference haplotype in that path, is a rich source of information about the historical relationships between the individuals and populations that the target and reference haplotypes originate from.

The connection of this analysis to admixture is both simple and subtle. If the target haplotype is from an admixed individual, and the reference haplotypes represent the admixture sources, the reconstructed path along the target haplotype will approximate the true underlying ancestry (Hellenthal et al., 2014), sometimes called ancestry tracts (e.g. Gravel, 2012). Difficulty arises because, in practice, the reference haplotypes do not represent a single ancestry, due to incomplete lineage sorting or gene flow. This means that it is naive to assume that matching a reference haplotype sampled from a certain population represents ancestry from that population.

One way to attempt to account for the complex historical relationships among reference haplotypes and the populations/ancestries they represent is by assigning the reference haplotypes to distinct groups and quantifying the degree of haplotype matching between those groups. Each group can then be identified by its distinct signature of haplotype matching to all other groups. The ancestry in the target can then be modeled as a mixture of these reference groups. This approach was taken in Lawson et al. (2012) and Chacón-Duque et al. (2018), and is the approach we used in manuscript 2.

1.2.2 Motivation and goals for manuscript 2

Manuscript 2 is motivated by the observation that European ancestry, in addition to Inuit ancestry, is an important aspect of the genetic makeup of the present day Greenlandic people. Despite this, we do not have a good understanding of how the history of Inuit/European contact has generated the current gene pool in Greenland. The Danish-Norwegian and later Danish colonial period and subsequent years as part of the Kingdom of Denmark have certainly had a large impact on Greenland, and likely also in the genetic composition of the Greenlandic people. But the extended and varied history of contact between the Greenlandic people and Europeans is broader than this relationship. Quantifying the genetic contributions of different European countries to Greenland can further enrich our understanding of the histories of both Greenlandic and European peoples.

Historical documents can shed some light on what to expect. There are records of marriages between Greenlanders and Danes, Norwegians, and Swedes going back to the 1740's, as well as census documents that estimated 8% of Greenlanders to have both Inuit and European ancestry around 1800 (Seiding, 2013). But existing records do not allow a full picture of gene flow from Europe. Some written accounts from visitors to Greenland are also available e.g. "... 9644 eskimos of which 3/4 has Danish blood in them" by Charles Francis Hall, polar explorer about the population of the Sisimiut region in Greenland (Hall, 1864). The Dutch may also be source of European ancestry in Greenland: "No wonder half of Sisimiut is said to be of Dutch descent." (Erngaard and Vejen, 1972), a reference to the whaling period in the 17th and 18th centuries when the Dutch sent up to 100 ships each year to Greenlandic waters and ports (Frandsen et al., 2017).

In manuscript 2 we sought to exploit the advent of the new powerful haplotype based methods to investigate to what extent different European countries have contributed to the genetic makeup of the present-day Greenlandic people. Prior to the study we expected that Denmark was a major source, but that also other countries, including the Netherlands and Norway were among the sources based on their history of contact with Greenland.

1.2.3 Results

To pursue our goal, we combined SNP data from 1582 admixed Greenlanders, 181 unadmixed Greenlanders, and 8275 Europeans from 14 countries to conduct a haplotype-based analysis of European ancestry sources in Greenland.

We then attempted to quantify the ancestry contributions from each of the 14 European countries in two ways, with group-based and individual-based analyses. In the group-based analysis, we estimated that Denmark contributed 91% of the European ancestry in Greenland, with the only other European country contributing more than 1% ancestry being Norway at 2.1%. This result suggests that Denmark has been the primary source of European ancestry in Greenland, substantially higher than other countries. In the individual-based analysis we estimated that 69.5% of the admixed Greenlanders had at least 5% Danish ancestry, many more than any other European reference country. In both analyses, we saw little evidence of ancestry from the British-Irish Isles, or from the Netherlands/Belgium, both regions with extended historical contact with Greenland.

We also performed an analysis to investigate the timing of admixture in Greenland, specially gene flow in the last few generations. Based on an analysis of local ancestry (Maples et al., 2013) within each admixed Greenlander, we identified a large number of individuals with local ancestry patterns consistent with very recent ancestors having 100% European ancestry. In total, we estimated up to 35% of the total European ancestry in the admixed Greenlanders to be consistent with a European ancestor in the previous generation. Many other individuals had local ancestry patterns consistent with a European ancestor in the previous three generations, suggesting much of the European ancestry in Greenland is very recent.

Taken together, these results suggested that most of the European ancestry is more recent than the start of the Danish colonial period, with little evidence of earlier European gene flow to Greenland not associated with Nordic countries. Specifically, the lack of any evidence of Dutch ancestry was surprising, as accounts such as Erngaard and Vejen (1972) suggest otherwise. However, as we discuss in the manuscript there may be reasonable explanation for this, including a severe epidemic in the area where the Dutch whalers stayed. And notably the ancestry source results fit with our observation that much of the gene flow appear to be recent, consistent with demographic statistics in Greenland and matching a time when Denmark was the primary European contact with Greenland.

1.2.4 My contributions

For manuscript 2, I performed all analyses, made all of the figures, and contributed to study design. Manuscript writing was shared between me, Aviaja Lyberth Hauptmann and Ida Moltke, with help

from co authors and Inge Høst Seiding, an expert on the recent history of Greenland as well as Anders Albrechtsen. Aviaja and Inge were instrumental in providing a historical context to the results of the genetic analyses. They also contributed to Figure 1. In a future version of this manuscript Aviaja may become joint first author.

1.2.5 Conclusions and future directions

In the manuscript, we address the analyses and their historical context in great detail. Below I will briefly discuss a few issues that arose during this study, as well as mention some opportunity to further this work.

Initially, we were worried that the relatively low number of overlapping SNP sites (135K) between the Greenlandic and European data sets would be a barrier to our ability to recover fine-scale population structure necessary for the ancestry analysis. However, it turned out to be sufficient for our goals. We were likely aided by the large sample sizes for the European reference countries (a total of 8275 individuals across the 14 countries). Early analyses with more SNPs, and fewer individuals per country had difficulty distinguishing between countries with similar ancestries such as Denmark vs Norway (data not shown). Analyses on a finer scale than countries such as in Leslie et al. (2015), may have higher data requirements.

The investigations into the timing of European gene flow into Greenland highlighted some of the complexities of dealing with recent and ongoing admixture. This type of admixture is not well approximated by pulse admixture models that are behind popular methods for inferring admixture history, often based on linkage disequilibrium (e.g. Loh et al., 2013; Hellenthal et al., 2014). Methods like this can be applied to the Greenlandic data, but it is difficult to interpret the results as the mean age of admixture is not very informative in this case. It is also possible to use the lengths of local ancestry tracts to date admixture events (e.g. Gravel, 2012), but we found it difficult to statistically phase the Greenlandic data with a low enough switch error rate to characterize the long local ancestry tracts that result from recent admixture. Plots of local ancestry along the genome of first generation offspring of one Inuit and one European parent suggested switch error rates that were sufficiently high to impede the identification of long ancestry tracts.

Instead, we used what we called the "ternary ancestry fraction" plots (see Figure 5, manuscript 2). These are a relatively simple way to summarise local ancestry fractions and address the phasing issues discussed above, as these fractions are robust to phasing switch errors. These summaries of local ancestry do seem to capture some important patterns of recent admixture in Greenland, and could be useful in other studies of recent admixture, especially if developed further. For example a similar idea was developed in Xue et al. (2017) to fit a pulse admixture model, but the ongoing and recent admixture in Greenland again complicates many useful assumptions when modeling admixture, such as independence of ancestry tracts on either side of an recombination event.

While the genetic history of Inuit and European contact in Greenland is certainly a topic of some interest, national and individual identity can be a delicate subject. This is especially the case when genetic results provide information about an individual's ancestors. We did not report individual-level ancestry results for privacy reasons. Recognising this, we are committed to communicating these results in a responsible

manner.

Indeed, this work has already benefited from outreach in Greenland. An early version of these results were publicly presented at the the Greenland National Museum & Archives in 2018 by co author Aviaja Lyberth Hauptmann. This lead to two large improvements in the study. First, our European reference countries did not include the Netherlands at that time, and she heard feedback from Greenlanders that the Dutch were important to have represented among the European references. We were able to include them in subsequent analyses. Second, this presentation played a vital role in the inclusion of Inge Høst Seiding, the head of archives at the Greenland National Museum, an invaluable addition. There are further plans to make this research relevant outside the academic world.

1.3 Estimating linkage disequilibrium in admixed populations

1.3.1 Background

Linkage disequilibrium

Linkage disequilibrium (LD) is a broad term used to describe the nonrandom association of alleles at different loci. LD is affected by many different evolutionary process, such as genetic drift, selection, recombination and mutation. Due to the effect of these genetic processes on LD, it is informative for building genetic maps (e.g. Myers et al., 2005), inferring natural selection Voight et al. (2006), as well as making inferences about demographic processes including population size changes Tenesa et al. (2007) and admixture Loh et al. (2013). An understanding of LD is also very important for the design of genetic studies, it can help predict the number of sites you need to cover the genome, and the power of genetic association studies Pritchard and Przeworski (2001). Furthermore, LD can help prune a set of loci so that statistical methods that treat each site independently can be reasonably applied.

LD is often measured across pairs of loci and there are number of different two-locus measures of LD that are sensitive to different aspects allelic associations. Below are the most common measures of LD for two loci, defined in terms of haplotype frequencies:

$$D_{AB} = p_{AB} - p_A p_B \tag{1.3}$$

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \tag{1.4}$$

$$D' = |\frac{D}{D_{max}}|. \tag{1.5}$$

Where A and B are alleles at different loci, p_x is the frequency of allele x, p_{AB} is the frequency of the haplotype carrying both the A and B alleles, and D_{max}, is the smaller of $p_A(1 - p_B)$ and $p_B(1 - p_A)$. These various measures of LD are all useful in different contexts, but r^2 is now likely the most commonly used measure as it has an interpretation in terms of the statistical Pritchard and Przeworski (2001) and genealogical independence of the two loci McVean (2002). One important note is that these measures are not independent of the two allele frequencies at the pair of loci; this complicates comparing LD measures between pairs of loci with different allele frequencies Hedrick (1987). As can be seen from the above equations, both D' and r^2 are adjusted by the allele frequencies at the two loci, in part so that values for pairs of loci with different allele frequencies are made more comparable .

Given a set of haplotypes or haplotype frequencies, the above LD measures can be calculated directly. However, many common methods of acquiring genotype data, such as next-generation sequencing or SNP arrays, do not intrinsically provide phase information across sites in diploid individuals, especially for loci separated by more than a few hundred base pairs. This means that for many types of available genetic data, the LD measures cannot be calculated directly, but need to be estimated. One option is to infer haplotypes from genotypes using statistical phasing (e.g. Stephens et al., 2001; Delaneau et al., 2011), an approach that leverages haplotype sharing across many individuals to help infer haplotype phase. But accurate statistical phasing is often reliant on utilizing sets of known reference haplotypes, and is therefore not always of sufficient quality. Another option is to use the genotypes to estimate

LD. There are multiple methods for the estimation of LD from genotype data, including methods that estimate haplotype frequencies from genotypes using an expectation-maximization algorithm (Excoffier and Slatkin, 1995) or solving a cubic equation (Gaunt et al., 2007), as well as other popular methods (Rogers and Huff, 2009).

In practice, the LD measures are applied to a sample of individuals or haplotypes that represent a larger group or population. A natural next step is to use the LD measured in this sample to estimate LD in the population. This is appealing because LD in populations can be related to both the recent and long-term effective population size of the population (Hayes et al., 2003; Waples, 2006; Tenesa et al., 2007).

But the extrapolation from sample LD to population LD is difficult. The sampling process itself produces LD, akin to the genetic drift that occurs each generation. Assuming we aim to estimate the LD in a given population from a random sample of size n, then the estimate will be biased and the size of the bias is dependant both on the sample size (approximately $1/n$ for loci not in LD), (Weir and Hill, 1980)), as well as the population-level LD. There are no unbiased estimators of r^2 (Ragsdale and Gravel, 2019), however there are methods that attempt to correct for this bias for pairs of loci with low (Weir and Hill, 1980) or high (Bulik-Sullivan et al., 2015) levels of LD, and addressing this bias is still an active area of research (e.g. Ragsdale and Gravel, 2019).

1.3.2 Linkage Disequilibrium decay curves

Recombination breaks up existing haplotypes when they pass from one generation to the next. In this way, recombination can reduce LD as alleles get shuffled onto new haplotypes each generation. Pairs of loci separated by greater genetic distances are more likely to have recombination events between them, and have lower expected values of LD. This leads to a natural summary of LD across many pairs of loci in the form of an LD decay curve. LD decay curves summarize the mean LD across a range of genetic distances and are presented as the primary LD summary statistic in many studies of humans or other species (e.g. 1000 Genomes Project Consortium et al., 2015; Franssen et al., 2015; Alves et al., 2019). Like LD generally, LD decay curves contain information about demographic processes, like population size changes. For example, LD for pairs of loci at close distances are more dependent on long-term effective population size, and LD for pairs of loci at further distances more dependent on recent population history (Hill, 1981; Hill and Weir, 1988). Most often decay of r^2 is presented, but other measures of LD can also be used (e.g. Abecasis et al., 2001). Different human populations have different LD decay curves, due to their different population histories. To illustrate this, I have included a figure of LD decay across the 26 population samples present in the 1000 Genomes Project (Figure 1.3A). This figure was part of the manuscript announcing the completion of the 1000G project 1000 Genomes Project Consortium et al. (2015). Evident in Figure 1.3A is the large variation in r^2 decay curves across the 1000G population samples. The most noticeable pattern is the difference between the LD curves for the non-African population samples (non-orange colors) and those for African population samples red-orange-yellow colors), with African populations having lower LD and a less steep decay, reflecting the distinct population histories of populations with an out-of-Africa bottleneck, and those without. Notice that the issue of sample size bias has been considered here by downsampling each population sample down to the same number of individuals ($n = 61$) making them comparable. In Figure 1.3B, I illustrate the effect of sample size on r^2 and r^2 decay curves; it contains seven LD decay curves from a single

population, measured across a range of sample sizes from $n = 5$ to $n = 679$. The upward bias in r^2 due to sample size is visually evident across the range of example sample sizes.

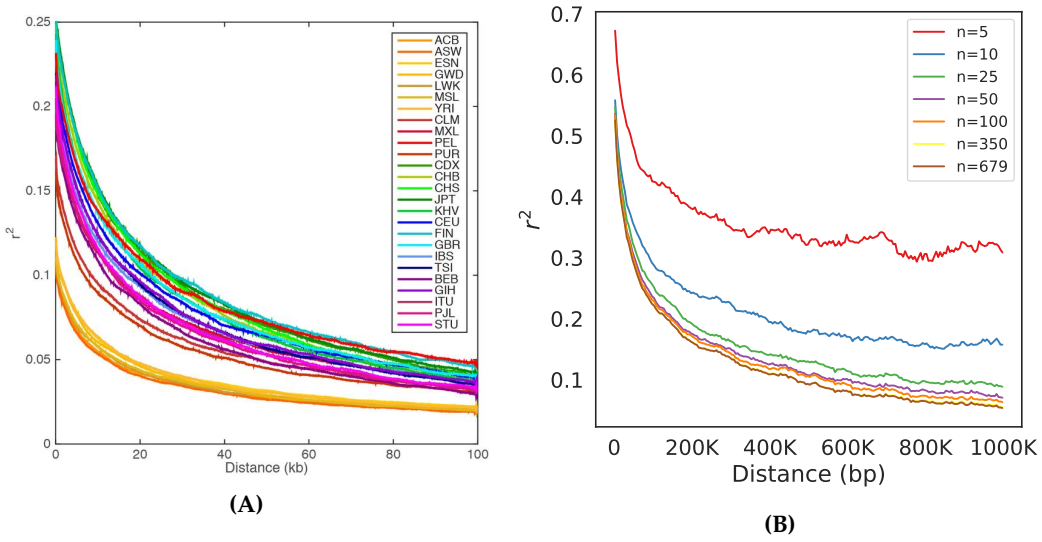


Figure 1.3: LD decay curves for human populations. **A**) LD decay curves across 10,000 randomly selected polymorphic sites in each population. Each population sample is thinned to n=61. The plotted line represents a 5 kb moving average, and the graph extends out to 100kb. **B**) Seven LD decay curves from unadmixed Greenlanders, thinned to samples sizes of [5, ... ,679], and only measured at sites with minor allele frequency > 0.05. The plotted line represents a 5kb binned average and the graph extends out to 1Mb. A) is taken directly from from 1000 Genomes Project Consortium et al. (2015) and is in the public domain.

1.3.3 Motivation and goals for manuscript 3

As discussed above, LD patterns can be related to long term and recent effective population sizes. However models that relate LD and N_e (e.g. Sved, 1971; Hill, 1981) do not directly account for gene flow between diverged populations, an important evolutionary force in many species (Supple and Shapiro, 2018). This is a significant limitation, as genetic admixture between diverged populations can have a substantial effect on LD patterns. At pairs of loci separated by a short genetic distance, admixture results in an LD pattern that is intermediate between the LD in the source populations of the admixture (Chakraborty and Weiss, 1988), see also Figure 1 of manuscript 3 for an example. This means that is is difficult to interpret LD decay curves in admixed populations, as they reflect both admixture and demographic history.

At pairs of loci separated by a large genetic distance, admixture can produce LD that is far above the LD in the source populations prior to admixture. LD decay curves at long distances (up to tens of millions of base pairs) in admixed populations can be used to infer the timing of admixture events. For example Loh et al. (2013) presents a method to fit a date of admixture to an observed pattern of long-range LD decay. To do so, they attempt to isolate the effects of admixture on LD by excluding LD that is present in the source populations of the admixture.

The goal of the third manuscript is in some sense the opposite: to provide a method that makes it

possible to isolate the LD that was present in the source populations of an admixed population by removing the effects of admixture on LD. Figure 1.4 demonstrates the basic idea.

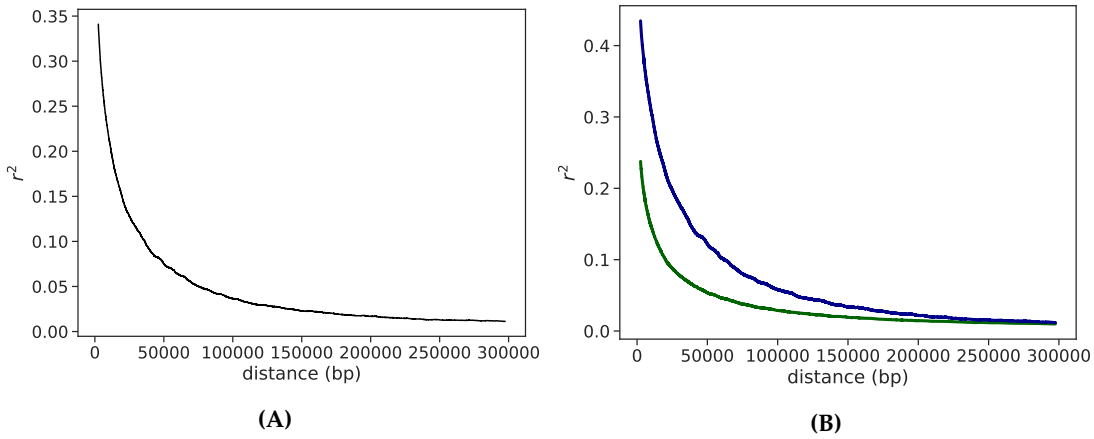
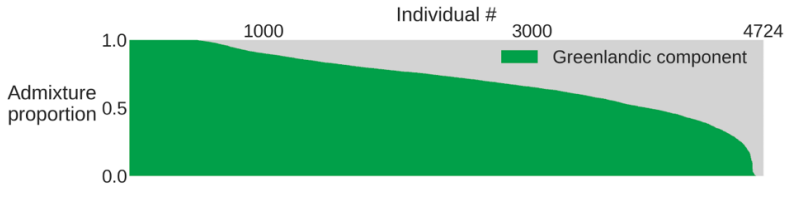


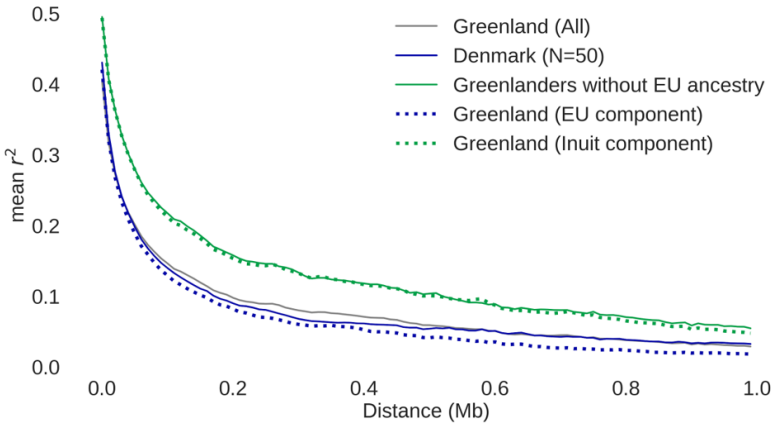
Figure 1.4: Diagram of a goal of LDadmix **A)** shows an example LD decay curve in a two-way admixed population, plotted out to 300Kb. **B)** shows the LD decay curve in each of the two admixture sources, blue and green. Notice the change of scale.

A method to do this is not needed if one has access to many individuals from each of the source populations, because in that case LD can just be estimated from directly those individuals. However, we were motivated by the observation that there are many human populations that are recently admixed to the extent that very few - if any unadmixed samples are available, making estimation of LD in at least one of the source populations difficult.

A method that fulfills this goal was introduced in Moltke et al. (2015) and applied to genotype data from a large sample of Greenlanders. The Greenlandic population is admixed with the majority of its ancestry from Inuit and the remaining ancestry is majority European (see Manuscript 2 for details). The estimated LD decay curves are shown in Figure 1.5, replotted by me from the results presented in that paper. Figure 1.5 shows five LD decay curves estimated from the genotype data. The gray line shows LD decay measured across all 4724 Individuals (4674 Greenlanders and 50 Danes). The solid blue line shows LD decay measured across the 50 Danes, a proxy for the European ancestry in the admixed Greenlanders. And the solid green line shows LD decay measured across a subset of the Greenlanders that only have Inuit ancestry. Notice the solid green LD decay curve is above and distinct from the solid blue and gray lines consistent with a long severe population bottleneck for the Greenlandic Inuit ancestry Pedersen et al. (2017). The dotted blue and green lines are the estimated LD decay curves for each ancestry (Inuit, European) as estimated by the method in the manuscript 2. The estimated curves are close to the LD decay curves estimated from unadmixed samples from each of the source populations of the Greenlandic population, with an especially close match for the Inuit ancestry.



(A)



(B)

Figure 1.5: A) Admixture bar plot of 4724 individuals (4674 Greenlanders and 50 Danes). The Inuit ancestry is shown in green. **B**) Five LD decay curves estimated from the individuals represented in the admixture plot. The gray line shows LD decay measured across all 4724 individuals. The solid blue line shows LD decay measured across the 50 Danes, a proxy for European ancestry. The solid green line shows LD decay measured across a subset of the Greenlanders that only have Inuit ancestry. The dotted blue and green lines are the estimated LD decay curves for each ancestry (Inuit, European) as estimated by the method introduced in the manuscript.

This result was encouraging, but left some open questions. Specifically, the method and result was simply presented as-is with no evaluation of the method's performance more generally, e.g. via application to simulated data or any other data sets. Also, the number of individuals in this study was very large (>4.5K), many more than most studies will have access to, so it is unclear how the results are affected when fewer samples are available. In addition, many unadmixed Greenlanders and Danes were included in the analysis, so again it is unclear how much the results relied on their presence. Finally, the method was never implemented in a user friendly software package, making it difficult for others to apply. With this in mind we sought to:

- Evaluate the performance on simulated data with known LD
- Evaluate the performance with smaller sample sizes
- Evaluate to what extent the performance depends on the presence of entirely unadmixed samples - and the distribution of admixture proportions more generally
- Facilitate other researchers conducting studies of LD in admixed populations, by producing software to conduct the analyses

1.3.4 Results

The results from this work are presented in manuscript 3, it presents three major results: 1) a software program LDadmix, 2) evaluation of the performance of LDadmix on simulated data, 3) and application of LDadmix to admixed human populations from the Americas. LDadmix is an open-source software program that implements the method described in Moltke et al. (2015). It uses common genotype data formats, and reports estimates of haplotype frequencies as well as r^2, D, and D′ for each ancestry using genotype data from a group admixed individuals. It also requires external estimates of admixture proportions from a program like STRUCTURE (Pritchard et al., 2000) or ADMIXTURE (Alexander et al., 2009). LDadmix is implemented in Python 3.6 and utilizes multiprocessing, and shared memory access, together with and just-in-time compilation with Numba (Lam et al., 2015) to allow analysis of many pairs of loci in parallel with limited memory requirements.

Next, we presented results from testing the performance of LDadmix on simulated data. We showed that LDadmix can work well in a two-way admixture with as few as 200 individuals, many fewer than previously demonstrated. In addition, we showed how the accuracy of the LD estimates obtained with LDadmix depend on the distribution of admixture proportions across individuals. Both the sum as well as the variance in admixture proportions across individuals affected the ability of LDadmix to recover accurate estimates of LD. On the group of 200 individuals, LDadmix was robust to the lack of unadmixed samples, but had difficulty in cases with very little variance in admixture proportions across individuals.

Finally, we also presented results from applying LDadmix to data from admixed human populations from the Americas. Specifically, we applied it to data from the 1000 Genomes Project with a goal of reconstructing the LD decay pattern in the ancestral American population that contributed ancestry to four admixed American population samples of the 1000 Genomes dataset. We recovered an elevated LD curve for the American ancestry component, consistent with the demographic history of this ancestral population. We further applied LDadmix to the individual population samples, but found we had limited ability to recover population-specific LD patterns for the four populations, likely due to the limited amount of Native American ancestry in each population sample.

1.3.5 My contributions

I wrote the manuscript with help from Ida Moltke and Anders Albrechtsen. I conducted all analyses and simulations and made all figures. I authored the software program LDadmix, based on original code written for analysis of data in Moltke et al. (2015) by Anders Albrechtsen. The original idea for the model behind LDadmix was presented in Moltke et al. (2015), and I was not involved in that project.

1.3.6 Conclusions and future directions

Currently, researchers investigating short-range LD in admixed populations face a difficult choice. The LD in a random sample of individuals from the population is likely to be significantly impacted by admixture. But, depending on the admixture history of the population, non-admixed individuals may be rare. In these cases, excluding admixed individuals may reduce sample size, impacting estimates of

LD. LDadmix presents a solution to this problem, allowing the inclusion of admixed individuals, while accounting for the effect of admixture on LD.

Despite the promising results presented in the manuscript, there is plenty of room to more fully develop LDadmix as an approach to estimating LD in admixed populations. Below I will briefly discuss three topics that are possible points of improvement: 1) further assessment of LDadmix with simulated data, 2) investigating the 'effective sample size' of each ancestry, and 3) moving beyond estimation of r^2 decay curves.

There are number of ways that we could utilize further simulations to assess the performance of LDadmix in a wider range of scenarios. I think the most basic outstanding question is "how many individuals are needed?". While we present some guidance in the manuscript, this is difficult to answer in a simple manner as the correct answer depends on multiple factors, including the distribution of admixture proportions across individuals, the degree of genetic divergence between the admixing populations, and the amount of LD. A further analysis of how the accuracy of estimates of LD from LDadmix scale with sample size could help researchers by providing guidance in how to interpret LDadmix results in a number of different contexts.

In the manuscript, we evaluate LDadmix on simulated two-way (K=2) admixture scenarios. We also apply LDadmix to a real three-way (K=3) admixture and argue that the results we achieve are reasonable. In our analysis, accurate recovery of three distinct LD decay curves was likely possible because we had access to unadmixed individuals to serve as proxies for two out of the three ancestry sources. However, we did not evaluate LDadmix on a simulated $K = 3$ admixture scenario to demonstrate this directly, doing so would provide further confidence in our results.

The simulated data were constructed in such a way that post-admixture recombination within two-locus haplotypes was not not possible. This matches an assumption of LDadmix that two-locus haplotypes in the analyzed individuals are inherited without recombination from the source populations. This is equivalent to assuming that the two sites of each two-locus haplotype share a single source ancestry. In the case of recent admixture, recombination is unlikely to occur between loci at close genetic distances, so this is a very reasonable assumption, at least on average. That said, the manuscript currently does not investigate the effect of violating the assumption of no recombination. For inferring LD at short genetic distances in a population with recent admixture, I expect the effect to be small, but recombination places a limit on the usefulness of LDadmix for inferring LD for at pairs of loci separated by a large genetic distance, or in cases of non-recent admixture.

Sample sizes are very relevant for studies of LD. However, in a sample of individuals from an admixed population, it is not likely that all ancestries present in the population will be equally represented. This can be due to random sampling, or it can occur because the ancestries are not all equally frequent in the admixed population. In addition, if the admixture event is not too old there can be substantial variation in admixture proportions across individuals (Verdu and Rosenberg, 2011). We found that both the total sum of each ancestry, as well as its distribution across individuals affected the ability of LDadmix to correctly infer LD.

We attempted a few different methods to address this bias. A simple estimate of the sample size within each ancestry is available as the sum of the admixture proportions of that ancestry across individuals, but this turns out to be too optimistic, as it fails to account for the uncertainty in the source ancestry of

each haplotype. This uncertainty occurs because we are probabilistically modeling the ancestry source of each haplotype based on the admixture proportion of the individual it occurs in. To address this further, we tried to devise a notion of the "effective sample size" of each ancestry. We estimated the effective sample size of each ancestry based on the horizontal asymptote of the r^2 decay curve as genetic distance between loci increased. In practice, the asymptotic mean r^2 ($\hat{r}^2$) was measured for loci >10Mb apart, and the effective haploid sample size (n_{eff}) for each ancestry was estimated as: $n_{eff} = \frac{2}{\hat{r}^2}$.

While this notion of effective sample size accounted for at least some of the bias, it failed to capture some important aspects of the ancestry-specific LD patterns inferred by LDadmix. This suggests that a single effective sample size for each ancestry may not be appropriate across all pairs of loci.

Finally, in this manuscript we focused on one possible application of LDadmix; estimating ancestry-specific r^2 between pairs of loci across a range of distances, from a sample of admixed individuals, and using these estimates to generate an LD decay curve for each source ancestry. It is also possible to investigate LD for individual pairs of loci in more depth, including evaluating the precision of our two-locus estimates of LD. This could be useful to investigate ancestry-specific pattern of LD at sites of interest. LDadmix treats each pair of loci independently and so provides estimates of two-locus haplotype frequencies within each source ancestry. The manuscript shows that RMSD values for r^2 are relatively constant across a range of distances, although they can vary by ancestry. Further work to investigate the accuracy of two-locus haplotype frequencies in detail is warranted.

In conclusion, manuscript 3 develops and presents a novel way to analyze LD in admixed populations and by providing a software implementation, we aim to aid other researchers interested in LD in admixed populations. LDadmix is easy to use, especially compared to alternative methods of assessing LD decay in admixed populations that involve local ancestry assignment. In the manuscript, we have applied LDadmix to human data in this study, but it should be directly applicable to admixed individuals of any diploid species.

1.4 Concluding remarks

The work presented in this thesis touches on three important concepts in population genetics: relatedness, admixture, and linkage disequilibrium. While these concepts are not new, the work presented here, especially the first and third manuscripts, clearly demonstrates that there is still a need for new statistical methods addressing these topics. This is especially important when the simplifying assumptions made by many current methods, such as a lack of inbreeding or admixture do not hold true, as they do not for many species and populations. Almost everywhere where we look in human history, we find genetic evidence of admixture. Accounting for this and other evolutionary complexities will hopefully be a more common route for methods in the future.

If this evolutionary complexity is to be better understood, the ever-growing amount of genetic data produced can certainly help us, as long as it is properly modeled. The Greenlandic project presented in the second manuscript provides a hint at the level of genetic resolution and understanding that is possible to achieve with these expanding data sets. There will soon be many more opportunities to study all manner of our history, as well as that of many other species. We should be ready to deal with that coming reality it armed with statistical methods that address life's complexities as well as its inherent structure. I hope that the work presented here will end up playing a small part in that.

Bibliography

1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

G R Abecasis, E Noguchi, A Heinzmann, J A Traherne, S Bhattacharyya, N I Leaves, G G Anderson, Y Zhang, N J Lench, A Carey, L R Cardon, M F Moffatt, and W O Cookson. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, 68(1):191–197, January 2001.

Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, September 2010.

David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664, September 2009.

Joel M Alves, Miguel Carneiro, Jade Y Cheng, Ana Lemos de Matos, Masmudur M Rahman, Liisa Loog, Paula F Campos, Nathan Wales, Anders Eriksson, Andrea Manica, Tanja Strive, Stephen C Graham, Sandra Afonso, Diana J Bell, Laura Belmont, Jonathan P Day, Susan J Fuller, Stéphane Marchandeau, William J Palmer, Guillaume Queney, Alison K Surridge, Filipe G Vieira, Grant McFadden, Rasmus Nielsen, M Thomas P Gilbert, Pedro J Esteves, Nuno Ferrand, and Francis M Jiggins. Parallel adaptation of rabbit populations to myxoma virus. *Science*, 363(6433):1319–1326, March 2019.

Carlos Eduardo G Amorim, Stefania Vai, Cosimo Posth, Alessandra Modi, István Koncz, Susanne Hakenbeck, Maria Cristina La Rocca, Balazs Mende, Dean Bobo, Walter Pohl, Luisella Pejrani Baricco, Elena Bedini, Paolo Francalacci, Caterina Giostra, Tivadar Vida, Daniel Winger, Uta von Freeden, Silvia Ghirotto, Martina Lari, Guido Barbujani, Johannes Krause, David Caramelli, Patrick J Geary, and Krishna R Veeramah. Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.*, 9(1):3547, September 2018.

Georgios Athanasiadis, Doug Speed, Mette K Andersen, Emil V R Appel, Niels Grarup, Ivan Brandslund, Marit Eika Jørgensen, Christina Viskum Lytken Larsen, Peter Bjerregaard, Torben Hansen, and Anders Albrechtsen. Estimating narrow-sense heritability from genome-wide data in admixed populations. preprint, March 2019.

Nathan A Baird, Paul D Etter, Tressa S Atwood, Mark C Currey, Anthony L Shiver, Zachary A Lewis, Eric U Selker, William A Cresko, and Eric A Johnson. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, 3(10):e3376, October 2008.

Elena Bosch, Francesc Calafell, Zoë H Rosser, Søren Nørby, Niels Lynnerup, Matthew E Hurles, and Mark A Jobling. High level of male-biased scandinavian admixture in greenlandic inuit shown by y-chromosomal analysis. *Hum. Genet.*, 112(4):353–363, April 2003.

Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, June 2013.

Sharon R Browning and Elizabeth A Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, April 2012.

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726): 203–209, October 2018.

Juan-Camilo Chacón-Duque, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuña-Alonzo, Rodrigo Barquera, Mirsha Quinto-Sánchez, Jorge Gómez-Valdés, Paola Everardo Martínez, Hugo Villamil-Ramírez, Tábita Hünemeier, Virginia Ramallo, Caio C Silva de Cerqueira, Malena Hurtado, Valeria Villegas, Vanessa Granja, Mercedes Villena, René Vásquez, Elena Llop, José R Sandoval, Alberto A Salazar-Granara, Maria-Laura Parolin, Karla Sandoval, Rosenda I Peñaloza-Espinosa, Hector Rangel-Villalobos, Cheryl A Winkler, William Klitz, Claudio Bravi, Julio Molina, Daniel Corach, Ramiro Barrantes, Verónica Gomes, Carlos Resende, Leonor Gusmão, Antonio Amorim, Yali Xue, Jean-Michel Dugoujon, Pedro Moral, Rolando González-José, Lavinia Schuler-Faccini, Francisco M Salzano, Maria-Cátira Bortolini, Samuel Canizales-Quinteros, Giovanni Poletti, Carla Gallo, Gabriel Bedoya, Francisco Rothhammer, David Balding, Garrett Hellenthal, and Andrés Ruiz-Linares. Latin americans show wide-spread converso ancestry and imprint of local native ancestry on physical appearance. *Nat. Commun.*, 9(1):5388, December 2018.

R Chakraborty and K M Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. U. S. A.*, 85(23):9119–9123, December 1988.

Yoonha Choi, Ellen M Wijsman, and Bruce S Weir. Case-control association testing in the presence of unknown relationships. *Genet. Epidemiol.*, 33(8):668–678, December 2009.

Matthew P Conomos, Alexander P Reiner, Bruce S Weir, and Timothy A Thornton. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.*, 98(1):127–148, January 2016.

Charles William Cotterman. *A calculus for statistico-genetics*. PhD thesis, The Ohio State University, 1940.

Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nat. Methods*, 9(2):179–181, December 2011.

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

Thomas Derrien, Jordi Estellé, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. Fast computation and applications of genome mappability. *PLoS One*, 7 (1):e30377, January 2012.

Erik Erngaard and Georg Vejen. *Grønland i tusinde år*. Sesam, 1972.

L Excoffier and M Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, 12(5):921–927, September 1995.

Niels Frandsen, Gulløv Hans H, Heinrich Jens C, Einar L Jensen, Ole Marquardt, Søren Rud, Inge Seiding, Peter A Toft, and Søren Thuesen. Grønland – den arktiske koloni. In Hans Christian Gulløv, editor, *Danmark og kolonierne*, pages 46–107. Gads Forlag, København, 2017.

Susanne U Franssen, Viola Nolte, Ray Tobler, and Christian Schlötterer. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental drosophila melanogaster populations. *Mol. Biol. Evol.*, 32(2):495–509, February 2015.

T Max Friesen and Charles D Arnold. The timing of the thule migration: New dates from the western canadian arctic. *Am. Antiq.*, 73(3):527–538, 2008.

Matteo Fumagalli, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, Pascale Gerbault, Line Skotte, Allan Linneberg, Cramer Christensen, Ivan Brandslund, Torben Jørgensen, Emilia Huerta-Sánchez, Erik B Schmidt, Oluf Pedersen, Torben Hansen, Anders Albrechtsen, and Rasmus Nielsen. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1347, September 2015.

Tom R Gaunt, Santiago Rodríguez, and Ian Nm Day. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. *BMC Bioinformatics*, 8:428, November 2007.

Kirsty A Golding, Ian A Simpson, J Edward Schofield, and Kevin J Edwards. Norse–Inuit interaction and landscape change in southern greenland? a geochronological, pedological, and palynological investigation. *Geoarchaeology*, 26(3):315–345, 2011.

Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191(2):607–619, June 2012.

B Grønnow and M Sørensen. *Dynamics of Northern Societies*. National Museum of Denmark, Copenhagen, May 2006.

Hans Christian Gulløv. The nature of contact between native greenlanders and norse. *Journal of the North Atlantic*, pages 16–24, July 2008.

Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, 19(2):318–326, February 2009.

Charles Francis Hall. *Life With The Esquimaux: The Narrative Of Captain Charles Francis Hall*. Sampson Low Son & Marston London, 1864.

O Handt, M Krings, R H Ward, and S Pääbo. The retrieval of ancient human DNA sequences. *Am. J. Hum. Genet.*, 59(2):368–376, August 1996.

Wei Hao, Minsun Song, and John D Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, March 2016.

Ben J Hayes, Peter M Visscher, Helen C McPartlan, and Mike E Goddard. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.*, 13(4):635–643, April 2003.

P W Hedrick. Gametic disequilibrium measures: proceed with caution. *Genetics*, 117(2):331–341, October 1987.

Garrett Hellenthal, George B J Busby, Gavin Band, James F Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, February 2014.

W G Hill and B S Weir. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.*, 33(1):54–78, February 1988.

W G Hill and B S Weir. Variation in actual relationship as a consequence of mendelian sampling and

linkage. *Genet. Res.*, 93(1):47–64, February 2011.

William G Hill. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.*, 38(3):209–216, December 1981.

A Jacquard. Genetic information given by a relative. *Biometrics*, 28(4):1101–1114, December 1972.

M Kardos, G Luikart, and F W Allendorf. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity*, 115(1):63–72, July 2015.

Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15:356, November 2014.

Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: a LLVM-based python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, page 7. ACM, November 2015.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genet.*, 8(1):e1002453, January 2012.

W-C Lee. Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. *Ann. Hum. Genet.*, 67(Pt 6):618–619, November 2003.

Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Daniel J Lawson, Daniel Falush, Colin Freeman, Matti Pirinen, Simon Myers, Mark Robinson, Peter Donnelly, and Walter Bodmer. The fine-scale genetic structure of the british population. *Nature*, 519(7543):309–314, March 2015.

Hong Li, Gustavo Glusman, Chad Huff, Juan Caballero, and Jared C Roach. Accurate and robust prediction of genetic relationship from whole-genome sequences. *PLoS One*, 9(2):e85437, February 2014.

Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, December 2003.

Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254, April 2013.

M Lynch and J B Walsh. Genetics and analysis of quantitative traits. sunderland, MA: Sinauer assocs. *Inc. 980p*, 1998.

Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, November 2010.

Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, 93(2):278–288, August 2013.

Alicia R Martin, Konrad J Karczewski, Sini Kerminen, Mitja I Kurki, Antti-Pekka Sarin, Mykyta Artomov, Johan G Eriksson, Tõnu Esko, Giulio Genovese, Aki S Havulinna, Jaakko Kaprio, Alexandra Konradi, László Korányi, Anna Kostareva, Minna Männikkö, Andres Metspalu, Markus Perola, Rashmi B Prasad, Olli Raitakari, Oxana Rotar, Veikko Salomaa, Leif Groop, Aarno Palotie, Benjamin M Neale, Samuli Ripatti, Matti Pirinen, and Mark J Daly. Haplotype sharing provides insights into Fine-Scale population history and disease in finland. *Am. J. Hum. Genet.*, 102(5):760–775, May 2018.

Gilean A T McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–991, October 2002.

Ida Moltke, Niels Grarup, Marit E Jørgensen, Peter Bjerregaard, Jonas T Treebak, Matteo Fumagalli, Thorfinn S Korneliussen, Marianne A Andersen, Thomas S Nielsen, Nikolaj T Krarup, Anette P Gjesing, Juleen R Zierath, Allan Linneberg, Xueli Wu, Guangqing Sun, Xin Jin, Jumana Al-Aama, Jun Wang, Knut Borch-Johnsen, Oluf Pedersen, Rasmus Nielsen, Anders Albrechtsen, and Torben Hansen. A common greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature*, 512(7513):190–193, August 2014.

Ida Moltke, Matteo Fumagalli, Thorfinn S Korneliussen, Jacob E Crawford, Peter Bjerregaard, Marit E Jørgensen, Niels Grarup, Hans Christian Gulløv, Allan Linneberg, Oluf Pedersen, Torben Hansen, Rasmus Nielsen, and Anders Albrechtsen. Uncovering the genetic history of the present-day greenlandic population. *Am. J. Hum. Genet.*, 96(1):54–69, January 2015.

Jose Manuel Monroy Kuhn, Mattias Jakobsson, and Torsten Günther. Estimating genetic kin relationships in prehistoric populations. *PLoS One*, 13(4):e0195491, April 2018.

Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, and Peter Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, October 2005.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.*, 91(5):809–822, November 2012.

Casper-Emil T Pedersen, Kirk E Lohmueller, Niels Grarup, Peter Bjerregaard, Torben Hansen, Hans R Siegismund, Ida Moltke, and Anders Albrechtsen. The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: Insights from the greenlandic inuit. *Genetics*, 205(2): 787–801, February 2017.

Vania Pereira, Carmen Tomas, Juan J Sanchez, Denise Syndercombe-Court, António Amorim, Leonor Gusmão, Maria João Prata, and Niels Morling. The peopling of greenland: further insights from the analysis of genetic diversity using autosomal and x-chromosomal markers. *Eur. J. Hum. Genet.*, 23(2): 245–251, February 2015.

J K Pritchard and M Przeworski. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, 69(1):1–14, July 2001.

J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

Andrea S Putnam and Jamie A Ivy. Kinship-based management strategies for captive breeding programs when pedigrees are unknown or uncertain. *J. Hered.*, 105(3):303–311, May 2014.

Maanasa Raghavan, Michael DeGiorgio, Anders Albrechtsen, Ida Moltke, Pontus Skoglund, Thorfinn S Korneliussen, Bjarne Grønnow, Martin Appelt, Hans Christian Gulløv, T Max Friesen, William Fitzhugh, Helena Malmström, Simon Rasmussen, Jesper Olsen, Linea Melchior, Benjamin T Fuller, Simon M Fahrni, Thomas Stafford, Jr, Vaughan Grimes, M A Priscilla Renouf, Jerome Cybulski, Niels Lynnerup, Marta Mirazon Lahr, Kate Britton, Rick Knecht, Jette Arneborg, Mait Metspalu, Omar E Cornejo, Anna-Sapfo Malaspinas, Yong Wang, Morten Rasmussen, Vibha Raghavan, Thomas V O Hansen, Elza Khusnutdinova, Tracey Pierre, Kirill Dneprovsky, Claus Andreasen, Hans Lange, M Geoffrey Hayes, Joan Coltrain, Victor A Spitsyn, Anders Götherström, Ludovic Orlando, Toomas Kivisild, Richard Villems, Michael H Crawford, Finn C Nielsen, Jørgen Dissing, Jan Heinemeier,

Morten Meldgaard, Carlos Bustamante, Dennis H O'Rourke, Mattias Jakobsson, M Thomas P Gilbert, Rasmus Nielsen, and Eske Willerslev. The genetic prehistory of the new world arctic. *Science*, 345 (6200):1255832, August 2014.

Aaron P Ragsdale and Simon Gravel. Unbiased estimation of linkage disequilibrium from unphased data. preprint, April 2019.

Morten Rasmussen, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, Ene Metspalu, Toomas Kivisild, Ramneek Gupta, Marcelo Bertalan, Kasper Nielsen, M Thomas P Gilbert, Yong Wang, Maanasa Raghavan, Paula F Campos, Hanne Munkholm Kamp, Andrew S Wilson, Andrew Gledhill, Silvana Tridico, Michael Bunce, Eline D Lorenzen, Jonas Binladen, Xiaosen Guo, Jing Zhao, Xiuqing Zhang, Hao Zhang, Zhuo Li, Minfeng Chen, Ludovic Orlando, Karsten Kristiansen, Mads Bak, Niels Tommerup, Christian Bendixen, Tracey L Pierre, Bjarne Grønnow, Morten Meldgaard, Claus Andreasen, Sardana A Fedorova, Ludmila P Osipova, Thomas F G Higham, Christopher Bronk Ramsey, Thomas V O Hansen, Finn C Nielsen, Michael H Crawford, Søren Brunak, Thomas Sicheritz-Pontén, Richard Villems, Rasmus Nielsen, Anders Krogh, Jun Wang, and Eske Willerslev. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282): 757–762, February 2010.

Alan R Rogers and Chad Huff. Linkage disequilibrium between loci with unknown phase. *Genetics*, 182 (3):839–844, July 2009.

Noah A Rosenberg. Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.*, 70(Pt 6):841–847, November 2006.

Aylwyn Scally. The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.*, 41:36–43, August 2016.

Inge Seiding. *Married to the Daughters of the Country: intermarriage and intimacy in Northwest Greenland ca.1750 to 1850*. PhD thesis, University of Greenland, 2013.

Vladimir Shchur and Rasmus Nielsen. On the number of siblings and p-th cousins in a large population sample. *J. Math. Biol.*, 77(5):1279–1298, November 2018.

Martin Sikora, Vladimir V Pitulko, Vitor C Sousa, Morten E Allentoft, Lasse Vinner, Simon Rasmussen, Ashot Margaryan, Peter de Barros Damgaard, Constanza de la Fuente Castro, Gabriel Renaud, Melinda Yang, Qiaomei Fu, Isabelle Dupanloup, Konstantinos Giampoudakis, David Bravo Nogues, Carsten Rahbek, Guus Kroonen, Michäel Peyrot, Hugh McColl, Sergey V Vasilyev, Elizaveta Veselovskaya, Margarita Gerasimova, Elena Y Pavlova, Vyacheslav G Chasnyk, Pavel A Nikolskiy, Pavel S Grebenyuk, Alexander Yu Fedorchenko, Alexander I Lebedintsev, Sergey B Slobodin, Boris A Malyarchuk, Rui Martiniano, Morten Meldgaard, Laura Arppe, Jukka U Palo, Tarja Sundell, Kristiina Mannermaa, Mikko Putkonen, Verner Alexandersen, Charlotte Primeau, Ripan Mahli, Karl-Göran Sjögren, Kristian Kristiansen, Anna Wessman, Antti Sajantila, Marta Mirazon Lahr, Richard Durbin, Rasmus Nielsen, David J Meltzer, Laurent Excoffier, and Eske Willerslev. The population history of northeastern siberia since the pleistocene. preprint, October 2018.

Mikkel Sørensen and Hans Christian Gulløv. The prehistory of inuit in northeast greenland. *Arctic Anthropol.*, 49(1):88–104, 2012.

Doug Speed and David J Balding. Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.*, 16(1):33–44, January 2015.

Leo Speidel, Marie Forest, Sinan Shi, and Simon Myers. A method for genome-wide genealogy estima-

tion for thousands of samples. preprint, February 2019.

Jeffrey Staples, Dandi Qiao, Michael H Cho, Edwin K Silverman, University of Washington Center for Mendelian Genomics, Deborah A Nickerson, and Jennifer E Below. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am. J. Hum. Genet.*, 95(5):553–564, November 2014.

M Stephens, N J Smith, and P Donnelly. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68(4):978–989, April 2001.

Eric L Stevens, Greg Heckenberg, Elisha D O Roberson, Joseph D Baugher, Thomas J Downey, and Jonathan Pevsner. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet.*, 7(9):e1002287, September 2011.

Megan A Supple and Beth Shapiro. Conservation of biodiversity in the genomics era. *Genome Biol.*, 19 (1):131, September 2018.

J A Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.*, 2(2):125–141, June 1971.

Albert Tenesa, Pau Navarro, Ben J Hayes, David L Duffy, Geraldine M Clarke, Mike E Goddard, and Peter M Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, 17(4):520–526, April 2007.

Christoph Theunert, Fernando Racimo, and Montgomery Slatkin. Joint estimation of relatedness coefficients and allele frequencies from ancient samples. *Genetics*, 206(2):1025–1035, June 2017.

E A Thompson. The estimation of pairwise relationships. *Ann. Hum. Genet.*, 39(2):173–188, October 1975.

Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, June 2013.

Lucy van Dorp, Sara Lowes, Jonathan L Weigel, Naser Ansari-Pour, Saioa López, Javier Mendoza-Revilla, James A Robinson, Joseph Henrich, Mark G Thomas, Nathan Nunn, and Garrett Hellenthal. Genetic legacy of state centralization in the kuba kingdom of the democratic republic of the congo. *Proc. Natl. Acad. Sci. U. S. A.*, 116(2):593–598, January 2019.

Paul Verdu and Noah A Rosenberg. A general mechanistic model for admixture histories of hybrid populations. *Genetics*, 189(4):1413–1426, December 2011.

Benjamin F Voight and Jonathan K Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.*, 1(3):e32, September 2005.

Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS Biol.*, 4(3):e72, March 2006.

J Wang. Estimating pairwise relatedness in a small sample of individuals. *Heredity*, 119(5):302–313, November 2017.

Jinliang Wang. Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol.*, 107:4–13, February 2016.

Robin S Waples. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*. *Conserv. Genet.*, 7(2):167, March 2006.

B S Weir and W G Hill. Effect of mating structure on variation in linkage disequilibrium. *Genetics*, 95(2): 477–488, June 1980.

Sewall Wright. Coefficients of inbreeding and relationship. *Am. Nat.*, 56(645):330–338, 1922.

James Xue, Todd Lencz, Ariel Darvasi, Itsik Pe'er, and Shai Carmi. The time and place of european admixture in ashkenazi jewish history. *PLoS Genet.*, 13(4):e1006644, April 2017.

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569, July 2010.

Paper I:
Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data

By

Ryan K. Waples[1], Anders Albrechtsen[1], Ida Moltke[1]

1 The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Publication details

Published in *Molecular Ecology* Vol. 28: 35-48 (2019)
doi.org/10.1111/mec.14954

Received: 21 March 2018 | Accepted: 12 October 2018

DOI: 10.1111/mec.14954

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data

Ryan K. Waples | Anders Albrechtsen | Ida Moltke

Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen N, Denmark

Correspondence
Anders Albrechtsen and Ida Moltke, Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen N, Denmark.
Emails: albrecht@binf.ku.dk; ida@binf.ku.dk

Funding information
RKW and IM were supported by a grant from the Independent Research Fund Denmark awarded to IM (DFF – 4090-00244). AA was supported by a grant from the Lundbeck Foundation (R215-2015-4174).

Abstract

Knowledge of how individuals are related is important in many areas of research, and numerous methods for inferring pairwise relatedness from genetic data have been developed. However, the majority of these methods were not developed for situations where data are limited. Specifically, most methods rely on the availability of population allele frequencies, the relative genomic position of variants and accurate genotype data. But in studies of non-model organisms or ancient samples, such data are not always available. Motivated by this, we present a new method for pairwise relatedness inference, which requires neither allele frequency information nor information on genomic position. Furthermore, it can be applied not only to accurate genotype data but also to low-depth sequencing data from which genotypes cannot be accurately called. We evaluate it using data from a range of human populations and show that it can be used to infer close familial relationships with a similar accuracy as a widely used method that relies on population allele frequencies. Additionally, we show that our method is robust to SNP ascertainment and applicable to low-depth sequencing data generated using different strategies, including resequencing and RADseq, which is important for application to a diverse range of populations and species.

KEYWORDS
ascertainment bias, IBD, identity by descent, low-depth, NGS, non-model, relatedness

1 | INTRODUCTION

The ability to infer the familial relationship between a pair of individuals from genetic data plays a key role in several research fields. In conservation biology, it is used to design breeding programmes that minimize inbreeding (Kardos, Luikart, & Allendorf, 2015), in archaeology it is helpful to understand burial patterns and other cultural traditions (Baca, Doan, Sobczyk, Stankovic, & Weglenski, 2012; Sikora et al., 2017), and in population and disease genetics it is often used to exclude relatives, because many analysis methods within those fields assume all analysed individuals are unrelated and violations of this assumption can lead to wrong conclusions (Balding, 2006).

Numerous pairwise relatedness inference methods have been developed, for example, Thompson (1975), Lee (2003), Purcell et al. (2007), Albrechtsen et al. (2009), Manichaikul et al. (2010), Stevens et al. (2011), Korneliussen and Moltke (2015), Conomos, Reiner, Weir, and Thornton (2016), Dou et al. (2017), and many are available in popular software packages, like PLINK (Purcell et al., 2007), and KING (Manichaikul et al., 2010). Most of these methods estimate either the three relatedness coefficients k_0, k_1 and k_2, or the kinship coefficient $\theta = \frac{k_1}{4} + \frac{k_2}{2}$ for each pair of diploid individuals, where k_0, k_1 and k_2 are the proportions of the genome where a pair of individuals share 0, 1 or 2 alleles identical by descent (IBD) (Thompson, 2000). By definition, alleles are IBD when they

are identical due to *recent* common ancestry, but because alleles can also be identical due to older common ancestry or recurrent mutations, IBD status cannot be directly observed. Therefore, the pairwise relatedness coefficients and the kinship coefficient have to be estimated, which can be done from patterns of observed genetic identity (identity by state; in short IBS). Once estimated, these relatedness statistics, k_0, k_1, k_2 and the kinship coefficient can be used to infer familial relationships by comparison with the expectation of the statistics for different familial relationships (Hill & Weir, 2011).

Although inference of relatedness is of wide interest, most existing methods are not immediately applicable in studies with limited data or genetic resources. First, most existing methods require the allele frequencies of the source population. For most studies in modern humans, this is not a problem. However, in studies of ancient humans or other species, accurate estimates of population allele frequencies are often not obtainable, because only a low number of samples are available. Second, several existing methods consider consecutive loci jointly and use sliding windows or hidden Markov models to leverage the non-independence of allele sharing along the genome between relatives (Albrechtsen et al., 2009; Gusev et al., 2009; Kuhn, Jakobsson, & Gunther, 2018; Stevens et al., 2011). These methods are powerful, but require information about the genomic position of variable sites, and for non-model organisms, high-quality reference genome assemblies are often not available. Third, other methods avoid allele frequencies, but rely on access to many samples to provide a necessary context for relationship classification (e.g., Abecasis, Cherny, Cookson, & Cardon, 2001). Finally, nearly all existing methods—both frequency-based and others—require genotype data. However, in sequencing studies, samples are often only sequenced to low depth due to cost and technical issues. This makes it infeasible to call genotypes accurately (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012), precluding the use of these methods. There are a few methods that estimate relatedness from low-depth sequencing data by utilizing genotype likelihoods (e.g., Korneliussen & Moltke, 2015), or by using imputed genotype dosages (Dou et al., 2017). However, these methods function by leveraging access to many samples to estimate allele frequencies or perform accurate genotype imputation and are therefore not designed to apply to data sets with a low number of samples.

Hence, most existing methods to infer close familial relationships are not immediately applicable in studies where data are limited, including many studies of non-model organisms and ancient samples. One of the few exceptions is a simple but elegant test for pairwise relatedness proposed in Lee (2003). This test relies entirely on the relative frequency of different genotype combinations within a pair of individuals and thus only requires genotype data from the two target individuals. While useful, this test does not provide any means to distinguish between different types of close familial relationships; it only provides a statistical test for a pair of individuals of the null hypothesis of them being unrelated.

There are only a few methods that can be used to distinguish between different types of familial relationships for a pair of individuals when neither allele frequencies nor information about the relative genomic position of sites is obtainable. One such method consists of plotting the proportion of the genomic sites in which the two individuals share both alleles IBS (which we will denote IBS2) vs. the proportion of the genomic sites in which they share zero alleles IBS (which we will denote IBS0). This method was used in Rosenberg (2006), where it was applied to the Human Genome Diversity Project (HGDP) data set. In the resulting scatter plot of the HGDP data (Figure 1 of Rosenberg (2006)), pairs of individuals with the same relationship category form distinct clusters so that it is possible to locate parent–offspring pairs, full-sibling pairs and to a lesser extent more distant relationships such as half-siblings/avuncular/grandparent–grandchild and first cousins.

Another such method is based in part on the KING-robust kinship estimator (Manichaikul et al., 2010). The KING-robust kinship estimator was developed to be robust to population structure, but in practice it has been shown to provide biased kinship estimates when applied to pairs of samples whose four chromosomes are not all from the same population (Conomos et al., 2016; Thornton et al., 2012). However, the KING-robust kinship estimator is directly applicable to samples from the same homogenous population even when allele frequencies are unknown. The reason for this is that, like the test suggested in Lee (2003), it relies only on the genotype combinations within the two target individuals and does not require knowledge about of allele frequencies. Importantly, Manichaikul et al. (2010) show it is possible to infer if a pair of individuals are parent–offspring, full-sibling, half-siblings/avuncular/grandparent–grandchild, first cousins, or unrelated, by jointly considering KING-robust kinship and the fraction of sites IBS0 using SNP array data without allele frequencies. For example, see Figure 3a in Manichaikul et al. (2010); a scatterplot of KING-robust kinship vs. the fraction of sites IBS0.

However, both the methods described above have two important limitations. First, like most other methods to estimate relatedness, they were developed for genotype data only. For example, the KING software (Manichaikul et al., 2010) implementing the KING-robust kinship estimator requires genotype data as input, which can be problematic for studies where only moderate or low-depth sequencing data are available and calling genotypes is consequently difficult. Second, both methods rely on estimates of the fraction of sites IBS0, which can be problematic because this fraction, as well as the fraction of sites IBS1 and IBS2, is highly sensitive to SNP ascertainment. This means that the results of the methods are platform-dependent and are likely to differ between different SNP arrays and especially between SNP array and sequencing data sets. In turn, this means that it can be difficult to distinguish between full-siblings and parent–offspring pairs using these methods.

Motivated by the outlined limitations to the existing methods, we present a method for relationship inference that–unlike most existing methods–relies neither on allele frequencies nor on information about the relative position of the variant sites, and which–unlike other frequency-free methods— is (1) applicable even to sequencing data of so low depth that accurate genotypes cannot be called from it and (2) robust to SNP ascertainment bias.

The new method is inspired by previous methods; it uses the KING-robust kinship estimator and a statistic R0, which is similar to the test statistic from the test for relatedness suggested by Lee (2003). However, the method is new in two important ways. First, besides relying on the two statistics, R0 and KING-robust kinship, it also relies on a third new statistic, R1. More specifically, the method consists of using two combinations of these three statistics, R1–R0 and R1-KING-robust kinship, to infer relationships, and it is this combination of statistics that makes the method robust to ascertainment bias. Second, while the new method is straightforward to apply to genotype data like other similar methods, we also present two computational approaches to estimate the three statistics directly from sequencing data that take the uncertainty of genotypes into account, allowing application to low-depth sequencing data.

In the following, we first fully describe the three statistics, R0, R1 and KING-robust kinship, how they can be estimated and other methodological details. Next, using simulated and publicly available SNP array data, we show that the new method provides similar accuracy and precision to the commonly used frequency-based method implemented in PLINK, when such data are available. Then, using sequence data from the 1,000 Genomes Project (The Genomes Project, 2015), we show that the three statistics can be estimated directly from sequencing data of low depth (~4×), here defined as depth insufficient for accurate genotype calling. Moreover, we show that the estimates obtained in this way are useful for inference of close familial relationships and that this is not the case for estimates obtained from genotypes called from the same data. Using different subsets of the same data, we also show that this new method, unlike previous similar methods, is robust to SNP ascertainment. Finally, we show that the method also provides useful results when applied to sequencing data down-sampled to approximate data generated using reduced-representation approaches, for example, restriction site-associated DNA sequencing (RADseq) and discuss some potential applications and limitations of the new method.

2 | METHODS AND MATERIALS

2.1 | The R0, R1 and KING-robust kinship statistics

The method for relationship inference we propose consists of estimating three statistics called R0, R1 and KING-robust kinship from genetic data and interpreting plots of R1 vs. R0 and R1 vs. KING-robust kinship.

We define the three statistics, R0, R1 and KING-robust kinship in terms of the genomewide IBS-sharing pattern of two individuals of interest. At any given diallelic site, a pair of individuals will carry one of nine possible genotype combinations; the nine possible combinations of the two individuals each carrying 0, 1 or 2 copies of a specific allele, for example, the ancestral allele. We can therefore fully characterize the genomewide IBS-sharing pattern of a pair of individuals by nine counts or proportions denoted: A, B, C, D, E, F, G, H and I (Figure 1a), similar to a two-dimensional site-frequency spectrum (SFS) across the two individuals. The R0 and R1 statistics are defined as simple functions of a subset of these nine values as shown in Figure 1b,c, and the KING-robust kinship statistic, originally defined by Manichaikul et al. (2010), can also be re-formulated as a function of these 9 values (Figure 1d).

The new method is motivated by several observations. First, the expected values of A–I vary depending on the familial relationship between the pair of individuals of interest. Consequently, so do functions of A–I, including R0, R1 and KING-robust kinship. Notably, there is no overlap between the joint expectation ranges of [R1, R0] and [R1, KING-robust kinship] for the four close relationship categories: full-siblings (FS), half-siblings/avuncular/grandparent–grandchild (HS), first cousins (C1) and unrelated (UR) and the range of expected values for parent–offspring (PO) only overlaps with those of FS in a single point (Figure 2, for derivations see supplementary text). Crucially, this is true regardless of the underlying allele frequency spectrum and holds for any pair of non-inbred individuals from the same homogenous population, making [R1, R0] and [R1, KING-robust kinship] potentially useful for distinguishing between these relationships. Second, while A–I, and thus R0, R1 and the KING-robust kinship estimator can be calculated from genotype data, they can also be estimated directly from next-generation sequencing (NGS) data based on the expected number of sites with each genotype combination (see below for details). This makes the method appropriate even when the sequencing depth is too low for accurate genotype calling (see below for methodological details). Third, regardless of the type of data that is available, R0, R1 and KING-robust kinship can be estimated without the need for population allele frequencies or information about the relative position of the genomics sites analysed. Finally, we expect the three statistics to be robust to SNP ascertainment because they are ratios computed from sites that are variable within the two samples and should thus be unaffected by the number of non-variable sites and because the (unknown) underlying frequency spectrum should only have a limited effect on these ratios.

2.1.1 | Estimation from sequencing data

The counts of the nine genotype combinations, A–I, and thus R0, R1 and KING-robust kinship for a pair of individual, can be estimated directly from NGS data via the use of genotype likelihoods calculated from aligned sequencing reads. Genotype likelihoods provide a means to account for the genotype uncertainly inherent to low-depth NGS data. We used two distinct, but similar, approaches to estimate these statistics from sequencing data that both build on this idea.

The first approach, which we denote the IBS-based approach, considers all ten possible genotypes at each diallelic site for each of the two individuals of interest and consists of a maximum-likelihood (ML) estimation of the counts of each of the 100 (10 × 10) possible genotype pairs (for details, see supplementary text). To perform the ML estimation, we used an expectation–maximization (EM) algorithm, which we have added to the ANGSD software package (Korneliussen et al., 2014) "IBS". After obtaining the estimate of the

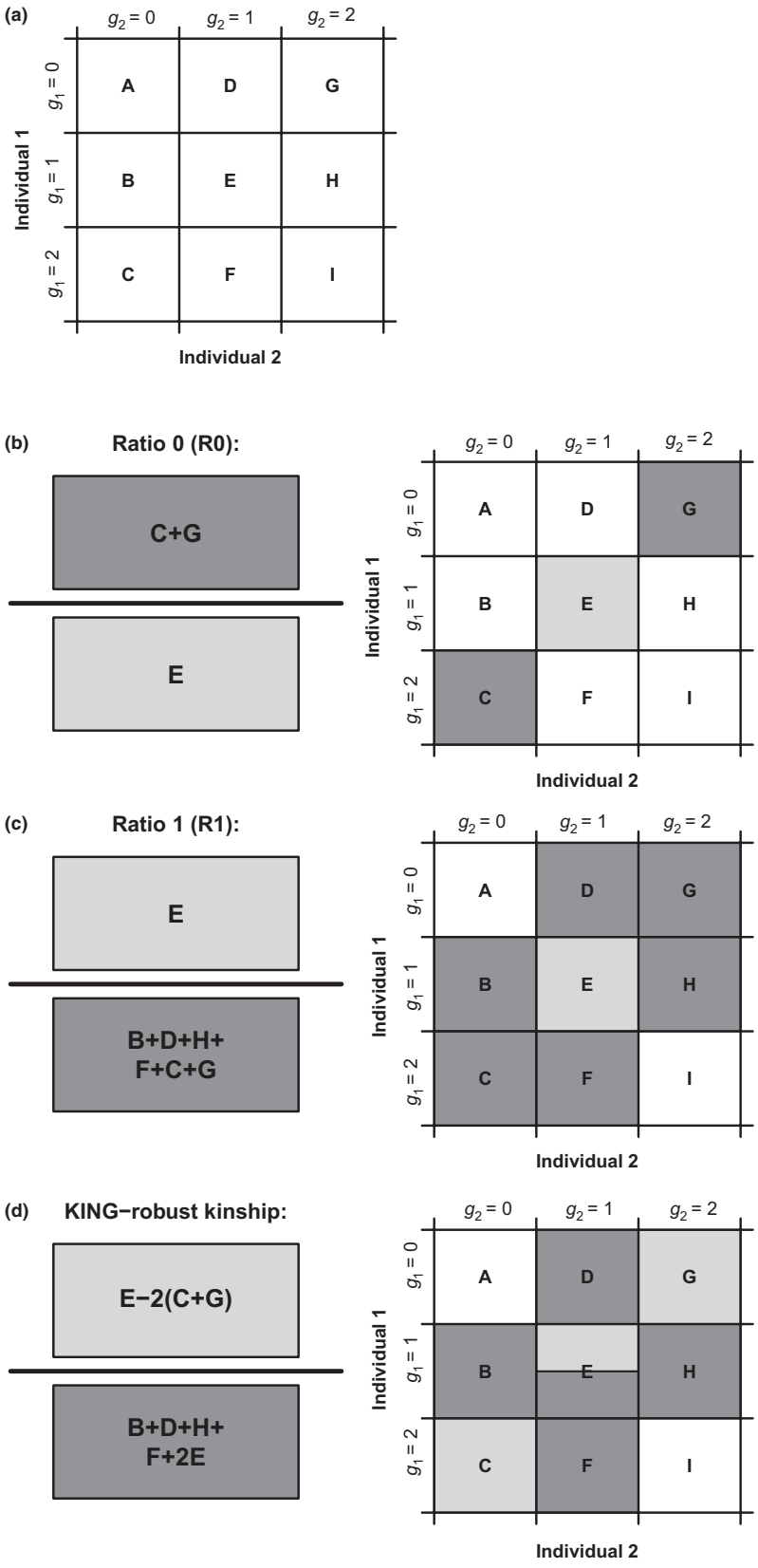


FIGURE 1 Definitions of pairwise genotype categories A–I and the R0, R1 and KING-robust kinship statistics. (a) Definition of the pairwise genotype categories A–I. Here, g_1 and g_2 denote the numbers of genotype for each of the two diploid individuals, 1 and 2, respectively. These genotypes are defined as the number of copies of a certain allele carried by individual 1 and 2, respectively. We assume diallelic variants such that g_1 and g_2 each has 3 possible values: 0, 1 and 2. For a pair of individuals, there are nine possible genotype combinations. We organize them into a 3 × 3 matrix and denote them with the letters from A to I. The values A–I can equivalently be either counts or proportions. (b) Definition of the R0 statistic based on the notation illustrated in (a). (c) Definition of the R1 statistic based on the notation illustrated in (a). (d) Definition of the KING-robust kinship estimator (Manichaikul et al., 2010), formulated using the notation illustrated in (a)

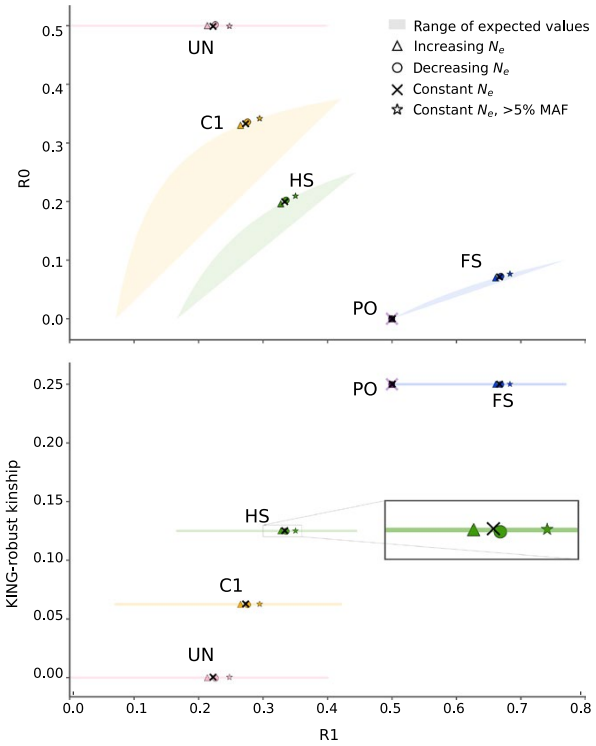


FIGURE 2 Ranges of expected values and simulation results for R1–R0 and for R1-KING-robust kinship for each of five relationship categories: parent–offspring (PO), full-siblings (FS), half-siblings/avuncular/grandparent–grandchild (HS), first cousins (C1) and unrelated (UR). (Top) The coloured shaded areas (sometimes just lines) show theoretically derived ranges of the joint expectation for R1–R0 based on expected IBD sharing (i.e., values of k_0, k_1 and k_2) for each relationship across all possible allele frequency spectra. For PO, this range is a singular point and is shown as a shaded purple X. The coloured symbols (triangle, circle, star) and black "X"s show values for each relationship obtained from data simulated under four different scenarios. Three of the scenarios are different demographic histories: (1) a 10-fold increase in *Ne* over the last 100 generations, (2) constant N_e and (3) a 10-fold decrease in N_e over the last 100 generations. The fourth scenario is also constant N_e, but sites are ascertained to have allele frequency above 5%. Note that, while it is difficult to see due to overplotting, all simulated values for PO fall very close to (R1, R0) = (0.5, 0). (Bottom) same as (Top), but for (R1, KING-robust kinship). Here all simulated values for PO fall very close to (R1, KING-robust kinship) = (0.5, 0.25)

counts of all the 100 possible genotype pairs, we converted them into estimates of A–I, by summing over the counts that correspond to each combination. For example, the genotype pairs: AA/AA, CC/CC, GG/GG and TT/TT all contribute to cells A or I of Figure 1. Counts corresponding to genotype pairs with more than two different alleles (e.g., AC/AG) were discarded. The advantage of this IBS-based approach is that it does not require specification of a known allele at each site and can thus be applied to nearly any sequencing data set, even with low-depth, without any prior information on the alleles at each site.

The second approach, which we denote the SFS-based approach, consists of performing ML estimation of the two-dimensional site-frequency spectrum (SFS) as in realSFS (Nielsen et al., 2012). To find the ML estimate of the SFS, we used an EM method implemented in the ANGSD software package under the name "realSFS" (Korneliussen, Albrechtsen, & Nielsen, 2014). The SFS-based approach requires one allele to be specified for each site, for example, the ancestral, the consensus or the reference allele. The model underlying this approach assumes that genotypes for each site have the possibility of containing this specified allele and up to one other unspecified allele. For the analyses performed in this paper, we used the consensus sequences from the highest depth individual (NA19042) to specify the alleles that exist at each site and restricted our analysis to sites where the depth in this individual was at least three.

The computational burden of analysing genomewide data sets can be significant. For both the genotype likelihood-based approaches described above, the main limitation is RAM, as data likelihoods for each site need to be loaded into memory for optimization. To overcome this limitation, we analysed each chromosome separately and then combined the values for each chromosome to produce a genomewide estimate. To calculate genotype likelihoods of the sequencing data, we used the original GATK genotype likelihood model (Mckenna et al., 2010) with independent errors, as implemented in ANGSD. We also tried to use the samtools genotype likelihood model (Li, 2011) with a more complicated error structure, but found it produced worse results (data not shown). Both IBS and realSFS produce the expected values for A–I, which we subsequently use to calculate R0, R1 and KING-robust kinship. For example command lines for each analysis, see supplemental text.

2.1.2 | Confidence intervals

All of the above estimation methods treat each site as independent. This assumption should not affect our expectation of each statistic (Wiuf, 2006), but statistical non-independence (here due to linkage disequilibrium (LD) and IBD) does affect standard estimates of uncertainty. To quantify the uncertainty, we therefore estimated confidence intervals for all statistics using a block-jackknife procedure. Confidence intervals were estimated by leaving each chromosome out (chromosome jackknife), which takes both the IBD and the LD correlation into account. The weighted block-jackknife variance estimator of Busing, Meijer, and Leeden (1999) was used to estimate the variance from the distribution of estimates for each statistic. The square root of this variance was interpreted as the standard error in our estimate.

2.2 | Application to simulated data

To evaluate the new method and to investigate the effects of demography on the expected statistics, we simulated genotype data under three different demographic histories: (1) a demography with constant effective population size (N_e), (2) a shrinking demography

with a 10-fold decrease in N_e over the last 100 generations, (3) an expanding demography with a 10-fold increase in N_e over the last 100 generations. For each of the three scenarios, we used the coalescent simulator msprime (Kelleher, Etheridge, & McVean, 2016) to simulate four haplotypes. These haplotypes were then used to construct pairs of related individuals in five relationship categories: parent–offspring (PO), full-siblings (FS), half-siblings/avuncular/grandparent–grandchild (HS), first cousin (C1) and unrelated (UR). For unrelated pairs of individuals, the genotype data were constructed by simply splitting the four haplotypes into two pairs. For related pairs of individuals, we constructed the genotypes at each variable site by first sampling whether the individuals shared 0, 1 or 2 alleles IBD according to the expected values of $[k_0, k_1, k_2]$ for the relevant relationship (Supplementary Table S1). Then, we used alleles present on the four haplotypes according to the sampled IBD status to construct the genotypes at that site. For example, if the individuals were sampled to share two alleles IBD at a site, both individuals were assigned the genotype consisting of the alleles present on the first two haplotypes at that site. The IBD sharing pattern was sampled independently for each SNP and thus LD and biological variation in IBD was not modelled. We concatenated the data from many independent simulations to achieve enough data so the IBD sharing was approximately equal to the expected values. Simulation code is available in the supplemental materials.

2.3 | Application to real data sets

To assess the utility of the new method on more realistic data, we applied it to two different publicly available data sets: SNP array data from seven HGDP populations (Rosenberg, 2006), and sequencing data from five related individuals from the Luhya in Webuye, Kenya (LWK) population of the 1,000 Genomes Project phase 3 (The Genomes Project 2015).

2.3.1 | HGDP SNP array

The HGDP SNP array data set was accessed 13 January 2017, and we followed the quality control steps described in Rosenberg (2006) to exclude mislabelled and duplicate samples. We selected seven populations from the HGDP based on the presence of several close familial relationships; five non-African populations: Surui, Pima, Karitiana, Maya and Melanesian, and two African populations: Mbuti Pygmies and Biaka Pygmies.

To ensure a fair comparison to the allele frequency-based inference method in PLINK, and by proxy to other commonly used methods, we constructed data sets where these methods have been shown to perform well. Specifically, we excluded individuals showing obvious signs of admixture (n = 16) or inbreeding (n = 2) from the selected HGDP populations. For details, see Supplemental text 2.3.1. This left us with a total of 142 individuals from the seven populations: Surui (n = 20), Pima (n = 20), Karitiana (n = 21), Maya (n = 16), Melanesian (n = 19), Biaka Pygmies (n = 31) and Mbuti Pygmies (n = 15). For each of these seven populations, we constructed a final set of genotypes by retaining genotypes from autosomal loci with genotyping rate >99%, minor allele frequency (MAF) >5%, Hardy-Weinberg equilibrium p-value > 10^{-4}.

The Ro, R1 and KING-robust kinship statistics for each of the 2,902 within-population pairs of individuals were then calculated from all sites where both individuals had non-missing genotypes.

2.3.2 | 1,000 Genomes sequencing data

To get sequencing data from several different relationship categories, we selected five individuals from two families in the Phase 3 1,000 Genomes (1000G) Luhya in Webuye, Kenya (LWK) population: NA19027, NA19042, NA19313, NA19331, NA19334. Across the five individuals, there is one pair of half-siblings (NA19027 & NA19042), and a separate trio of related individuals with a pair of full-siblings (NA19331 & NA19334), one parent–offspring relationship (NA19313 & NA19331) and another unspecified second-degree relationship (NA19313 & NA19334), possibly avuncular (The Genomes Project 2015). These stated relationships leave six unrelated pairs among the five individuals.

For each pair of the five LWK individuals, we estimated the R0, R1 and KING-robust kinship statistics in five different ways: (1) and (2) by applying the two different sequencing-based approaches described above to the 1000G aligned sequence data files (~4× coverage bam files), (3) by simple genotype counting based on the phased and high-quality curated genotypes provided in the hg37 1000G VCF files, (4) by genotype counting based on the subset of sites in approach 3 that overlap with the Illumina 650Y sites for the HGDP data (to investigate ascertainment, see below) and (5) by calling genotypes from the same 1000G bam files in a basic manner meant to mimic data from a species with a reference genome but few other genetic resources and then simply counting from the called genotypes. For genotype calling, we used samtools mpileup (v1.3.1) to summarize the reads overlapping each position, and bcftools call (v1.3.1) to assign the most likely genotype at each position. We used mostly default settings; non-default flags to samtools specified skipping indel positions. Non-default flags to bcftools specified using the consensus caller. For all sequence-based analyses (1, 2 and 5), we only considered reads with a minimum phred-scaled quality score of 30 and bases with minimum phred-scaled quality score of 20 and we restricted our analyses to genomic regions with a GEM 75mer mappability of 1 (Derrien et al., 2012). Notably, all the methods and filters used here can be applied to any study, including studies with only small contigs, for example made up of RAD loci, making the results relevant beyond resequencing studies utilizing well-assembled genomes.

2.4 | Assessing the effect of SNP ascertainment

To evaluate the effect of SNP ascertainment using real data, we created a subset of the curated genotype data from the five 1,000 Genomes individuals. We selected the sites that overlap with the Illumina 650Y array that was used for the HGDP and estimated our three relatedness statistics. We compared the results from this subset of HGDP sites to

results for the full genotype data set and also to the sequence-based analyses. For an additional comparison, we also performed the same comparison for the methods presented in Rosenberg (2006) and Manichaikul et al. (2010) by constructing scatterplots by their methods.

We also investigated the effect of SNP ascertainment using the data simulated from a constant demography (see "Application to simulated data" for details about the simulations) and compared results for the full data set to results obtained by including only sites with a minor allele count >2 out of 40 chromosomes (MAF > 5%) in the analyses.

2.5 | Assessing the effect of a limited number of sites

To assess the usefulness of the new method on data sets with fewer genomic sites covered by sequencing reads, we constructed reduced size data sets, in a way that mimicked some aspects of reduced-representation sequencing approaches such as RADseq. To produce each reduced data set, we selected a specific number of 200-bp windows randomly from the mappable genomic regions and restricted our analysis to sites falling within them. We used 10 k 50 k, 100 k and 250 k windows, representing ~4× sequencing coverage on 2 M, 10 M, 20 M or 50 M sites, respectively. All other aspects of the analyses were the same, expect for that for these data sets, we applied the IBS- and SFS-approaches to the complete data set, rather than splitting by chromosome as we did for the full data set. We suggest analysing the complete data in single run if your computational resources allow it, as we noticed some upward bias in the estimated number of IBS0 sites when the smaller data sets were analysed separately by chromosome.

2.6 | Comparison to other methods

To get a categorization of relationships for the HGDP data set described above based on a standard, commonly used allele frequency-based method, we first applied the allele frequency-based relatedness estimation algorithm in PLINK (v1.9) (Chang, Chow, Tellier, Vattikuti, & Purcell, 2015) to the individuals from each population separately to estimate the genomewide IBD fractions k_0, k_1 and k_2. Next, we applied the relationship criteria proposed in table 1 of Manichaikul et al. (2010) to the obtained estimates: the estimated k values were combined into an estimate of the kinship coefficient $\theta = \frac{k_1}{4} + \frac{k_2}{2}$, and a relationship degree was assigned to each pair of individuals based on comparing the estimated kinship coefficient to the criteria in the table. Parent–offspring and full-siblings were differentiated based on k_2 values. This provided us with a categorization into five categories: PO, FS, HS, C1 and UR. To achieve additional resolution, we further divided the last category (UR) into two: unknown/distantly related (UK-DR) and unrelated (UN). We did this by simply extending the logic behind the criteria proposed above. Specifically, we set the kinship threshold between UK-DR and UR to $1/2^{13/2}$, which corresponds to including 4th- to 5th-degree relatives in the UK-DR category.

To assess the accuracy and precision of the new method for familial relationship classification within the HGDP data, we examined concordance with the PLINK-based relationship categorization described above. For this purpose, we assigned a relationship category to each pair of individuals in two ways: (1) using the statistics R0 and R1 and (2) using a combination of KING-robust kinship and R0. For the former, we characterized each possible relationship by a single [R1, R0] point generated from data simulated under a demography with a constant population size over time, detailed in the "Application to simulated data" section and assigned each pair of individuals the relationship of the closest point using a Euclidean distance measure. For the latter, we used the KING-robust kinship criteria from table 1 of Manichaikul et al. (2010) as above. Since this table has overlapping kinship ranges for the PO and FS categories, we used the R0 statistic to distinguish PO from FS relationships: Ignoring rare effects like germline mutations and genotyping errors the expected value for R0 for PO relatives is zero, while for FS the value is above 0, we used an ad hoc cut-off of 0.02.

To estimate the statistics for identifying related individuals proposed by Rosenberg (Rosenberg, 2006) and KING (Manichaikul et al., 2010), we note that the KING-robust kinship estimator can (as previously described) be calculated directly from the same nine counts, A–I and so can the fraction of sites IBS0 and IBS2:

$$\text{KING}-\text{robust kinship} = (E - 2(C+G))/(B+D+H+F+2E)$$

$$\text{Fraction IBS0} = (C+G)/(A+B+C+D+E+F+G+H+I)$$

$$\text{Fraction IBS2} = (A+E+I)/(A+B+C+D+E+F+G+H+I)$$

We used these formulas in all our comparisons because this allowed us to estimate these statistics not only from genotype data but also directly from sequence data in the same manner as for R0 and R1. However, we note that this is our approach to estimating those statistics, and that existing tools like KING only allow users to estimate the statistics from genotype data.

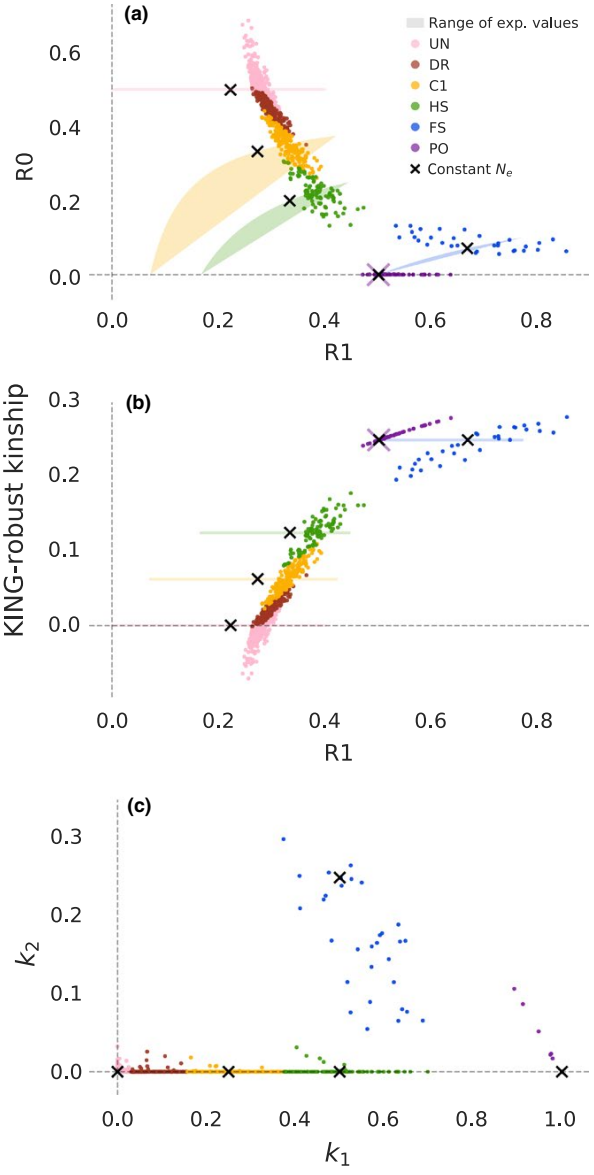
3 | RESULTS

To assess the performance of the new method, we first applied it to simulated genotype data to ensure that it works on sufficient data and to assess how sensitive it is to the underlying demographic history of the population the analysed samples are from. Next, we applied the new method to real data from different platforms to assess its performance on more realistic data. Finally, we performed a couple of additional analyses to access how robust the new method is to SNP ascertainment and to having data from only a limited number of sites available. Below we describe the results of all these analyses.

3.1 | Application to simulated data

We first applied the new method to simulated genotype data from several different relationship pairs from populations with three

different demographic histories: (1) constant N_e, (2) 10-fold increase in N_e over the past 100 generations, and (3) 10-fold decrease in N_e over the past 100 generations. We see very similar results across all three demographic scenarios, and in all cases, the R1–R0 and R1-KING-robust kinship values obtained from the simulated data were within the theoretically derived ranges of expectations (Figure 2). These results demonstrate that the method works if sufficient high-quality genotype data are available. Furthermore, they demonstrate that the range of examined population size histories has a limited effect, even though demographic history affects the allele frequency spectrum. In turn, this suggests that the range of expected values that are realistic for real data is markedly smaller than the theoretically possible ranges also shown in the figure, which is useful for classification purposes.



3.2 | Application to SNP array data

Next, we applied the method to SNP array data from the Human Genome Diversity Project (HGDP) to see how the method works on real data, for which standard allele frequency-based methods like PLINK are known to perform well. More specifically, we applied it to genotype data from unadmixed and non-inbred samples from seven populations originating from the HGDP. This resulted in the R1–R0 and R1-KING-robust kinship plots shown in Figure 3a,b (for population-specific plots, see Supplementary Figures S1 and S2). The true relationships for the pairs of individuals are not known; we instead coloured each point in Figure 3a,b according to the relationship category inferred based on results from the standard, commonly used allele frequency-based method PLINK (Figure 3c, Supplementary Figure S3). Since there are least 15 individuals in each of the seven selected HGDP populations, the allele frequency estimates for these populations should be reasonably accurate even with some relatedness. Hence, with the large amount of data available in this data set, the allele frequency-based method should provide correct inference of most, if not all, pairs closer than first cousins, but may not be able to fully distinguish first cousins from more distantly related pairs.

In Figure 3a,b, points from each relationship category clearly cluster together on both the R1–R0 plot and the R1-KING-robust kinship plot. Moreover, these clusters are located near both their theoretically derived ranges of expected values and the values from simulated data in a similar manner, the k_1–k_2 values for the same pairs of individuals cluster close to the expected and simulated values of k_1 and k_2 (Figure 3c). Almost all pairs identified as parent–offspring (PO) by the frequency-based method are easy to identify as such in both the R1–R0 plot and the R1-KING-robust kinship plot, which is not the case when only a single statistic is used (see also Supplemental Figures S1 and S2). The same is true for full-siblings (FS). Furthermore, points classified as half-siblings/avuncular/grandparent–grandchild (HS) or first cousins (C1) by the frequency-based method have a minimal overlap with each other and with less-related pairs (Figure 3a,b). The few

FIGURE 3 R1–R0 and R1-KING-robust kinship scatterplots for seven HGDP populations. Each coloured point represents a pair of individuals and is coloured according to the relationship category inferred using an allele frequency-based approach. Coloured shaded areas/lines show the theoretically derived range of expected values for specific relationship categories, as in Figure 2. Black "X"s show the values for a pair of individuals simulated under a constant population size, as in Figure 2. Note that in addition to the relationship categories for Figure 2 there is an additional category here representing distantly related pairs (DR). (a) R1–R0 plot for all pairs of individuals within each population (b) R1-KING-robust kinship plot for all pairs of individuals within each population. (c) Scatterplot of the two relatedness coefficients k_1 and k_2 for all pairs of individuals within each population estimated using the allele frequency-based approach implemented in PLINK. Note that the black "X"s here show simulated values for k_1 and k_2 and are not inferred by PLINK, they approximately coincide with the expected values of k_1 and k_2 for each relationship category (Supplementary Table S1)

pairs of individuals that were difficult to classify are the same pairs as those that are edge cases for the allele frequency-based method. This is apparent in an R1–R0 plot of the HGDP data constructed excluding pairs that are closer than 0.01 to the kinship coefficient thresholds that the frequency-based method used when classifying relationships (Supplementary Figure S4).

To quantify precision and accuracy, we examined the concordance between classifications based on the new method and the PLINK-based classification. We tried two simple classification schemes: one based on R1–R0, which uses proximity to the values we obtained from simulated data from a constant N_e demography, and one based on KING-robust kinship (for details see Methods and Materials). The results supported the visual assessment: both classification schemes are highly concordant with the classifications obtained using the frequency-based method (Supplemental Figure S5). Mean precision across all relationship categories was 0.90 for the R1–R0 method, vs. 0.89 for KING-robust kinship. Mean recall across all relationship categories was 0.88 for R1–R0, vs. 0.89 for KING-robust kinship. The relationship categories for which the method has the lowest precision are the first cousins vs. less-related pairs, where the allele frequency-based method is also known to have a hard time making classifications. For PO, FS and UR alone, the mean precision is as high as 0.99 for R1–R0 and 0.96 for KING-robust-kinship, and the mean recall for these three categories is as high as 0.96 for R1–R0 and 0.99 for KING-robust kinship. Hence, the new method provides comparable performance to a frequency-based method when sufficient genotype data are available, but without the need for allele frequency information.

3.3 | Application to sequencing data

To assess how well the new method works on more limited real data, we applied it to sequencing data from five low-depth (~4×) human genomes from the 1,000 Genomes project. Among the five selected samples, there is a parent–offspring pair, a pair of full-siblings, a pair of half-siblings, an unspecified 2nd-degree relationship (e.g., avuncular), and the rest are unrelated. We estimated the R0, R1 and KING-robust kinship for each pair in several ways. First, by using an IBS-based approach that estimates the proportion all pairwise combinations of the 10 possible genotypes (Figure 4, "IBS"). Second, by using an SFS-based approach where we estimated the two-dimensional site-frequency spectrum (2D-SFS) of each pair with a bi-allelic model and calculated R0, KING-robust kinship and R1 based on this spectrum (Figure 4, "realSFS"). Both these approaches base their estimates on genotype likelihoods calculated from the sequencing read data, instead of called genotypes, and take the uncertainty of the underlying genotypes that is inherent to low-depth sequencing data into account. The key difference between them is that the SFS-based approach requires specification of an allele known to exist at each site, whereas the IBS-based approach has no such requirement, making it more generally applicable. The approaches also differ in how they deal with sites with more than two unique alleles, either excluding them (IBS-based approach) or integrating over the two-allele possibilities (SFS-based approach), but these sites are rare (mean fraction as estimated by IBS: 1.8E-6) so the impact of discarding them is minimal.

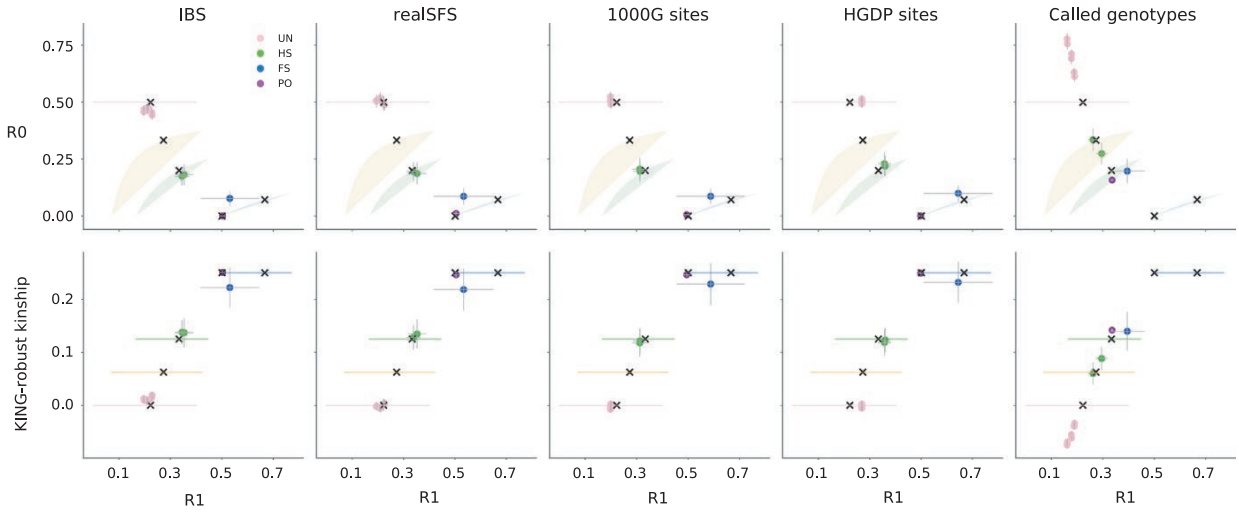


FIGURE 4 Relatedness plots for all pairs among five LWK individuals from the 1000G Project. (Top) R1–R0 scatterplots for pairs of five LWK individuals for five different analysis approaches: (1) IBS: estimation from (~4×) 1000G bam files, (2) realSFS: site-frequency spectrum-based estimation from (~4×) 1000G bam files, (3) 1000G sites: genotype counting using curated 1000G genotypes from the 1000G project, (4) HGDP sites: genotype counting using curated 1000G genotypes but only at sites that overlap with the Illumina 650Y array used for the HGDP, and (5) called genotypes: genotype counting using genotypes called de novo from (~4×) 1000G bam files. Points are coloured by their true relationship status, as reported by 1000G. Thin grey lines show confidence intervals (±2 SE) estimated using a chromosome jackknife. Coloured shaded areas/lines show the theoretically derived range of expected values for specific relationship categories from Figure 2. Black "X"s show the values for pairs with different relationships simulated under a constant population size, as in Figure 2. (Bottom) R1-KING-robust kinship scatter plots for the same data sets, confidence intervals and expected ranges are constructed in the same way

With the values estimated using the SFS-based approach, it is possible to visually classify of all the pairs to their relationship category within the set of close familial relationships (PO, FS, HS or UR), or by using one of the classification methods introduced earlier. Results for the IBS-based approach were similar, but unrelated individuals have a slight decrease in R0 and a slight increase in R1 and KING-robust kinship, compared to the SFS-based approach. This makes unrelated individuals appear slightly more related than expected for unrelated individuals from a homogenous population. However, despite this bias, it is still possible to correctly classify of all the pairs to their relationship category, suggesting that the IBS-based approach can be used when not enough information is available for the SFS-based approach. The chromosome block-jackknife estimates of uncertainty for the genotype likelihood-based methods were small, and varied by relationship type, with the pair of full-siblings having the most uncertainty in R0, R1 and KING-robust kinship.

We also calculated the three statistics from the high-quality phased genotypes for the same five individuals available from the 1,000 Genomes Project Phase 3 (Figure 4, "1000G sites") to see how well the two genotype likelihood-based approaches applied to low-depth sequencing data perform compared to direct calculations from high-quality genotype data for the same samples. In this comparison, results obtained by using the genotype likelihood-based approaches applied to low-depth sequencing data are close to those obtained from the high-quality genotypes for all the pairs (Figure 4).

Finally, we also made R1–R0 and R1-KING-robust kinship plots based on genotypes that we obtained through a standard genotype calling procedure from the raw read data. We did this to investigate whether the genotype likelihood-based approaches are necessary or one could just as well use genotypes called from the ~4× data. As expected, genotype calling had a large negative effect on the outcome; in the resulting R1–R0 and R1-KING-robust kinship plots, the half-siblings appear within the range of expected values for first cousins and both the parent–offspring and full-sibling pairs appear within the range of expected values for half-siblings (Figure 4, "called genotypes"). These results demonstrate the pitfalls of basing any relationship inferences, including R1–R0 and R1-KING-robust kinship plots, on genotypes called from low-depth data. Notably, this is also the case for the methods presented in Rosenberg (2006) and Manichaikul et al. (2010) (Figure 5, "called genotypes"). This clearly demonstrates that, with ~4× sequencing data, calling genotypes without external information, such as an imputation reference panel, is not a good alternative to a genotype likelihood-based approach. This implies that software packages designed to work only on genotype data, such as KING, should not be used on data like this.

3.4 | Assessing the effect of SNP ascertainment

To assess the effect of SNP ascertainment, we applied the new method to three different subsets of data from the five 1,000 Genomes

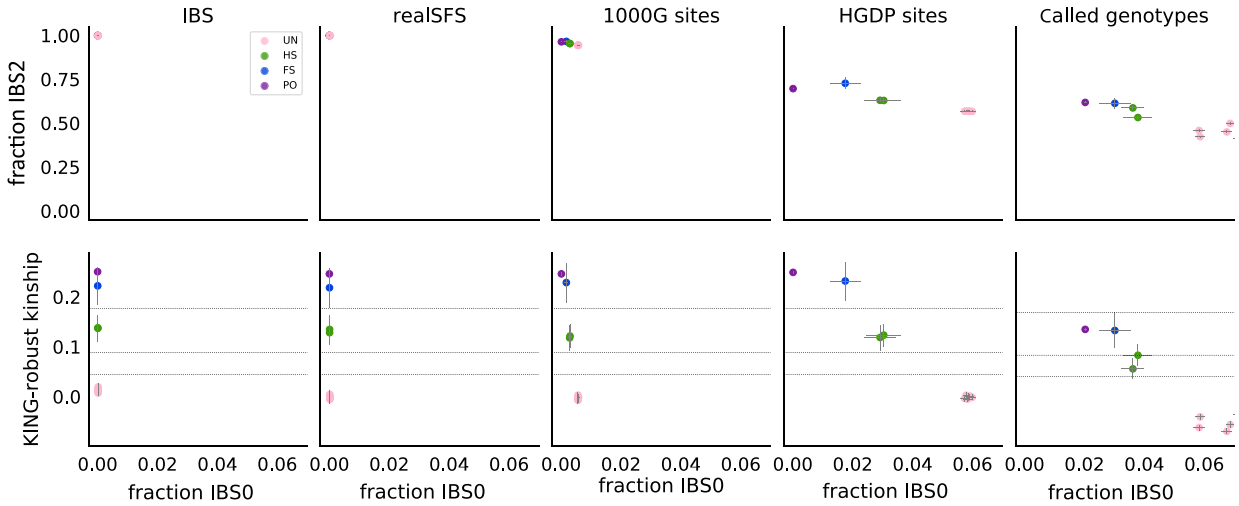


FIGURE 5 Results from two alternate frequency-free methods to different subsets and types of data from five 1000 Genomes samples. (Top) Results from applying the plotting approach from Rosenberg (2006) to pairs of the same five LWK individuals for five different analysis approaches: 1) IBS: estimation from (~4×) 1000G bam files, 2) realSFS: site-frequency spectrum based estimation from (~4×) 1000G bam files, 3) 1000G sites: genotype counting using curated 1000G genotypes from the 1000G project, 4) HGDP sites: genotype counting using curated 1000G genotypes at sites that overlap with the Illumina 650Y array used for the HGDP and 5) genotype counting using called genotypes: genotype called de novo from (~4×) 1000G bam files. Pairs are coloured by their true relationship status, as in Figure 3. Fraction IBS0/IBS2 are the overall fraction of sites that are IBS0/IBS2, res pectively. Grey lines centred on each point show confidence intervals (±2 SE) based on a chromosome jackknife. (Bottom) Results from applying the KING-robust based approach to the same pairs of LWK individuals using the same five different analysis methods as above. The horizontal black lines show the kinship thresholds used to distinguish unrelated (UR), first cousins (C1) half-siblings (HS), full-siblings (FS) and (PO) following (Manichaikul et al., 2010) from bottom to top, respectively. Thin grey lines centred on each point show confidence intervals (±2 SE) estimated using chromosome jackknife

individuals. The results for each of the three ascertainment schemes; all sites covered by sequencing data, 1,000 Genomes release sites and Illumina 650Y SNP array sites are similar (left four panels, Figure 4), showing SNP ascertainment does not have a large effect.

For comparison, we performed the same assessment of the methods presented in Rosenberg (2006) and Manichaikul et al. (2010) by constructing scatterplots of the same type as those shown in their papers (Figure 5). This revealed that both these other methods are much more affected by ascertainment than the method proposed here. In particular, the Rosenberg method is affected on both its *x*-axis (IBS0) and its *y*-axis (IBS2), which means that the expected region of the plot for each relationship will be different for different data sets (top row of Figure 5). The method presented in Manichaikul et al., 2010 is affected by the SNP ascertainment mainly on its *x*-axis (IBS0, bottom row of Figure 5). Therefore, the ascertainment mainly affects the ability to distinguish between parent–offspring and full-siblings, since the *y*-axis, which is only slightly affected by ascertainment, is the kinship coefficient, which can be used to distinguish between most close relationships except for parent–offspring and full-siblings. The *x*-axis, IBS0, is included in part to help make the distinction between PO and FS (Manichaikul et al., 2010), but this ability is clearly affected by SNP ascertainment (bottom row of Figure 5).

To further explore the effect of SNP ascertainment on the new method, we also performed analyses of the previously mentioned simulated of data from a population with a constant population size. This time we only analysed SNPs with MAF above 5% and compared the results to the results for the full data set. This confirmed the results from the real data analyses: SNP ascertainment does change the values a bit compared to when all sites are analysed, however the change is limited (Figure 2). This is well in line with the fact that we got very similar results for the simulated data from three populations with quite different population size histories and consequently different allele frequency spectra. Indeed, the effect of population size decline is similar to that of ascertaining for common SNPs, which makes sense because population decline is known to lead to a skew in the allele frequency spectrum towards more common SNPs.

3.5 | Assessing the effect of a limited number of sites

Genomewide shotgun sequencing data, as is available for the 1000G individuals, is not available for all species. Studies may instead have RADseq or similar data, covering only a fraction of genomic sites. To assess to what extent the new method can be used to analyse such data sets, we performed analyses of subsets of the 1000G data, constructed to mimic RAD sequencing data. Specifically, we analysed four subsets that consisted of 10 k, 50 k, 100 k and 250 k, 200 bp windows, representing 2 M, 10 M, 20 M or 50 M sites, respectively. For all but the smallest data subset, the point estimates were similar to those obtained using the full data set, showing the method is applicable when reducing the number of sites even with ~4× coverage (Figure 6, supplemental file 1). This suggests that even with the reduced number of sites tested, there was sufficient data to characterize the genomewide mean IBD fractions for both closely related and unrelated pairs. The uncertainty in the estimates, as estimated by a chromosome jackknife, increased with fewer sites, but the effect was limited, suggesting the biological variation in IBD sharing across chromosomes was larger than sampling variance across the examined sites.

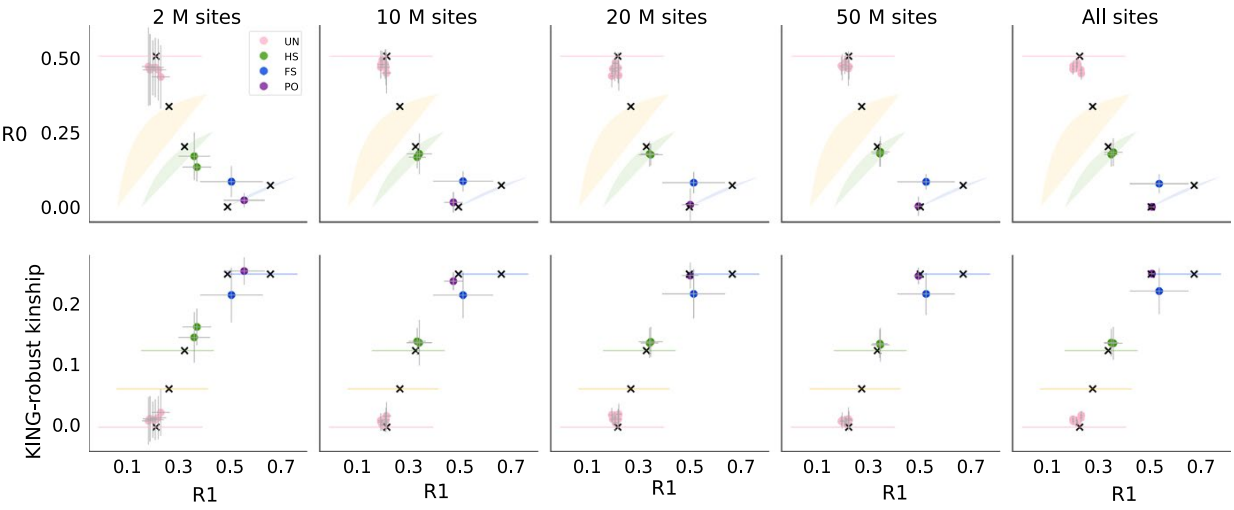


FIGURE 6 The effect on estimates of R0, R1 and KING-robust kinship from reducing the number of sites covered by sequencing data from the same five LWK individuals as in Figures 4 and 5. Each point shows the point estimate and error bars show ±2 SE estimated using chromosome jackknife. Each column shows results for different numbers of examined basepairs, including non-variable sites. Pairs of individuals are coloured by their true relationship. These plots show the results of the IBS-based method, see supplemental for results from the SFS-based approach. Coloured shaded areas/lines show the theoretically possible range of expected values for specific relationship categories from Figure 2 and black "X"s show the values for different relationship pairs of individuals simulated under a constant population size, as in Figure 2

4 | DISCUSSION

We have presented a simple new method for inferring if and how two individuals are related based solely on genetic data from the two target individuals. We demonstrated two ways in which it can also be applied directly to sequencing data via genotype likelihoods. And importantly, we showed that, the method provides useful results when applied to ~4× sequencing data as well as RADseq like subsets of such data. All of this combined implies that—unlike previous methods—this new method can be used even if all you have is low-depth sequencing data from a few individuals from a species without a reference genome.

4.1 | Comparison to similar methods

The new method is based on plotting two statistics, R0 and KING-robust kinship against a third, new statistic R1. The R0 statistic similar to the test statistic proposed by Lee (2003) to test for relatedness. The only differences are that the numerator and denominator are flipped and that E, the proportion of sites where both individuals are heterozygous, is included in the denominator in the statistic defined by Lee but absent in R0. R0 is also similar to the pairwise population concordance (PPC) statistic in PLINK (Purcell et al., 2007), a test of if the genotypes of a pair of individuals have more IBS0 sites than two unrelated individuals with the same ancestry are expected to have, signalling they have different ancestries. The KING-robust kinship estimator was proposed by Manichaikul et al. (2010) and implemented for genotype data in the program KING. Here, we extend it to estimation directly from sequencing data. Notably, our results suggest that this extension is vital for successful application to low-depth sequencing data, because estimates based on genotypes called from low-depth sequencing data are very poor (rightmost panels of Figures 4 and 5), which makes programs like KING inappropriate to apply to such data. This extension can also be used for similar pairwise statistics and thus makes existing methods based on such statistics, like KING-robust kinship, more widely applicable.

However, this extension is not the only contribution of this study. Another key new contribution is to provide an alternative to the IBS0 statistic (the proportion of sites where the two individuals share zero alleles IBS) that was utilized by Rosenberg (2006) and Manichaikul et al. (2010). As we have shown, the fraction of sites that are IBS0 or IBS2 is very sensitive SNP ascertainment, meaning that results are only comparable within each ascertainment scheme. The method from Rosenberg (2006), where IBS0 is combined with IBS2, is difficult, if not impossible, to use for relatedness inference in general because the fraction of sites that are IBS2 and IBS0 varies so wildly across different ascertainment schemes, such as between SNP arrays and sequencing data. On the other hand, KING-robust kinship is still very useful, but it loses the ability to distinguish between parent–offspring and full-siblings, as IBS0 was used for this. Due to this sensitivity to ascertainment, samples cannot be analysed in isolation and must be placed in the context of other samples with known relationships and the same ascertainment scheme. This requirement makes it difficult to apply these previous methods to ancient humans or other species with limited sample sizes.

In contrast, the ability to identify relatives based on expected values is maintained in the new method, regardless of ascertainment scheme due to the use of R0, instead of IBS0, which makes the new method robust to SNP ascertainment. Parent–offspring pairs tend to have an R0 estimate extremely close to 0, making them particularly easy to identify via the R1–R0 plot. The R1-KING-robust kinship plot, on the other hand, has the appealing aspect that the kinship axis has a biological interpretation, defined as the probability that two alleles sampled at random from two individuals are identical by descent. Hence, the two plots types, R1–R0 and R1-KING-robust kinship, each have their advantages. Finally, it is worth noticing that the two plots types seem to work better than a range of other plots constructed from similar ratio statistics that we explored (Supplementary Figure S6).

4.2 | Limitations and applications

While the new method provides substantial advantages over previous methods in situations with limited data, it does have some limitations. First, like most other relatedness inference methods, such as PLINK, the proposed method assumes that the individuals are not inbred and that they originate from the same homogenous population. And like many other relationship inference methods, it is not necessarily robust to violations of these assumptions. Previous studies have shown the effect of population structure and admixture on relatedness inference is complex and can potentially lead to bias in either direction depending on the circumstances, and this is true even for KING, which was developed to be robust to population structure (Conomos et al., 2016; Ramstetter et al., 2017; Thornton et al., 2012). Specific methods have been developed to correct for admixture when the allele frequencies in the admixing populations are known (e.g., Thornton et al., 2012; Moltke & Albrechtsen, 2014), or enough samples are available (Conomos et al., 2016; Dou et al., 2017). But since these methods work by exploiting knowledge about allele frequencies or access to many samples for their correction, the pairwise R0, R1 and KING-robust kinship statistics cannot be easily corrected in a similar manner. However, we note that Lee (2003) showed that the statistic he proposed for testing for relatedness can also be used to detect if two unrelated samples are not from the same homogeneous population. If this is the case, Lee's statistic will be significantly smaller than 2/3; and equivalently R0 will be significantly above 0.5, which may be helpful when interpreting R0, R1, KING-kinship plots in the presence of admixture or population structure more generally. Regarding inbreeding, one potential way to assess if one of the individuals is inbred is to compare heterozygosities across individuals; non-inbred and non-admixed individuals from the same population should have similar heterozygosity, so marked heterozygosity differences can be a warning signal.

A second limitation, which is shared with other relatedness estimation methods, is that there is significant biological variation in the

amount of IBD sharing between relatives with the same pedigree relationship due to randomness inherent in the process of recombination (Hill, 1993; Rasmuson, 1993). For humans, this means that a pair of relatives, say first cousins, will sometimes share less of their genomes IBD than another pair with a more distant pedigree relationship, say second cousins. This makes classification into specific relationships difficult. The degree of biological variation in IBD sharing between relatives varies across species and can even differ between sexes due to sex-specific recombination patterns. This makes it difficult to provide general guidance appropriate for all species. In general, species with more chromosomes and more recombination will have less variation in IBD sharing for a defined pedigree relationship, making it easier to distinguish among various potential relationship categories. To quantify this uncertainty, we propose a chromosomal bootstrap procedure that can be used if reads can be assigned to chromosomes.

Biological variation in IBD sharing is also related to the estimation and interpretation of confidence intervals on statistics like R0, R1 and KING-robust kinship. Relatedness and limited recombination also cause correlation between sites in the genome, due to shared IBD segments and LD. This correlation between sites increases the variance in the estimates of these statistics in a way that can be difficult to fully account for when computing confidence intervals. For statistics that test for introgression such as the D-statistic (Patterson et al. 2012), where the main concern is correlation due to LD, a block jackknife, leaving out contiguous blocks (e.g., 5 Mb) is a common approach. When considering relatedness, we want to compare our estimates to the expectations of each relationship category. Since shared IBD segments can be much longer than the range of LD we propose a more appropriate chromosome jackknife. In either case, a jackknife (or bootstrap) over single sites will fail to provide a confidence interval that accounts for the non-independence of the sites. For more discussion on this topic, see Thompson (2013). Unfortunately, this means that it is difficult to provide the most appropriate confidence intervals when no information about genomic positions is available.

Despite these limitations, we believe that the results presented here suggest the new method constitutes a helpful new tool for relatedness inference for studies with limited data. Identifying related samples is a crucial step in nearly any genetic analysis and can also reveal other problems such as duplicate samples or cross-contamination of genetic material. Removing the requirements to specify allele frequencies and to have accurate genotypes has the potential allow the identification of relatives even in small studies of non-model species or ancient samples. These types of studies do not currently have many good options to address relatedness.

AUTHOR CONTRIBUTIONS

AA, IM and RKW designed the research. RKW performed research, with input from IM and AA. IM and RKM wrote the paper together, with help from AA.

DATA ACCESSIBILITY

The IBS method is available at: http://www.popgen.dk/software/index.php/IBSrelate.

The data sets used are publicly available.

The HGDP SNP array data are available at: ftp.cephb.fr/hgdp_supp1.

The 1000G phase 3 aligned sequencing data are available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data.

The 1000G phase 3 called genotypes are available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

ORCID

Ryan K. Waples https://orcid.org/0000-0003-0526-6425
Anders Albrechtsen https://orcid.org/0000-0001-7306-031X
Ida Moltke https://orcid.org/0000-0001-7052-8554

REFERENCES

Abecasis, G. R., Cherny, S. S., Cookson, W., & Cardon, L. R. (2001). GRR: Graphical representation of relationship errors. *Bioinformatics*, *17*, 742–743. https://doi.org/10.1093/bioinformatics/17.8.742

Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F. C., & Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology*, *33*, 266–274. https://doi.org/10.1002/gepi.20378

Baca, M., Doan, K., Sobczyk, M., Stankovic, A., & Weglenski, P. (2012). Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community. *BMC Genetics*, *13*, 30. https://doi.org/10.1186/1471-2156-13-30

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*, 781–791. https://doi.org/10.1038/nrg1916

Busing, F. M. T. A., Meijer, E., & Van Der Leeden, R. (1999). Delete-m jackknife for unequal m. *Statistics and Computing*, *9*, 3–8.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics*, *98*, 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Derrien, T., Estelle, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. *PLoS ONE*, *7*, e30377. https://doi.org/10.1371/journal.pone.0030377

Dou, J., Sun, B., Sim, X., Hughes, J. D., Reilly, D. F., Tai, E. S. ... Wang, C. (2017). Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS Genetics*, *13*, e1007021. https://doi.org/10.1371/journal.pgen.1007021

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L. ... Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*, 318–326.

Hill, W. G. (1993). Variation in genetic identity within kinships. *Heredity*, *71*, 652–653. https://doi.org/10.1038/hdy.1993.190

Hill, W. G., & Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, *93*, 47–64. https://doi.org/10.1017/S0016672310000480

Kardos, M., Luikart, G., & Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity*, *115*, 63–72. https://doi.org/10.1038/hdy.2015.17

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*, e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*, 4009–4011. https://doi.org/10.1093/bioinformatics/btv509

Kuhn, J. M. M., Jakobsson, M., & Gunther, T. (2018). Estimating genetic kin relationships in prehistoric populations. *PLoS ONE*, *13*, e0195491.

Lee, W. C. (2003). Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. *Annals of Human Genetics*, *67*, 618–619. https://doi.org/10.1046/j.1529-8817.2003.00063.x

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Moltke, I., & Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics*, *30*, 1027–1028. https://doi.org/10.1093/bioinformatics/btt652

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, *7*, e37558. https://doi.org/10.1371/journal.pone.0037558

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J. … Williams, A. L. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, *207*, 75–82.

Rasmuson, M. (1993). Variation in genetic identity within kinships. *Heredity*, *70*, 266–268. https://doi.org/10.1038/hdy.1993.38

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, *70*, 841–847. https://doi.org/10.1111/j.1469-1809.2006.00285.x

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A. … Willerslev, E. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*, *358*, 659–662.

Stevens, E. L., Heckenberg, G., Roberson, E. D., Baugher, J. D., Downey, T. J., & Pevsner, J. (2011). Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genetics*, *7*, e1002287. https://doi.org/10.1371/journal.pgen.1002287

The Genomes Project. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics*, *39*, 173–188. https://doi.org/10.1111/j.1469-1809.1975.tb00120.x

Thompson, E. A. (2000). *Statistical inferences from genetic data on pedigrees NSF-CBMS regional conference series in probability and statistics (Vol. 6)*. Beachwood, OH: IMS.

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*, 301–326. https://doi.org/10.1534/genetics.112.148825

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *American Journal of Human Genetics*, *91*, 122–138. https://doi.org/10.1016/j.ajhg.2012.05.024

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *Journal of Mathematical Biology*, *53*, 821–841. https://doi.org/10.1007/s00285-006-0031-0

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Waples RK, Albrechtsen A, Moltke I. Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Mol Ecol*. 2019;28:35–48. https://doi.org/10.1111/mec.14954

Supplemental information for: Allele frequency-free inference of close familial relationships from genotypes or low depth sequencing data

Ryan K. Waples, Anders Albrechtsen, Ida Moltke

June 12, 2019

Contents

1 Supplemental Figures

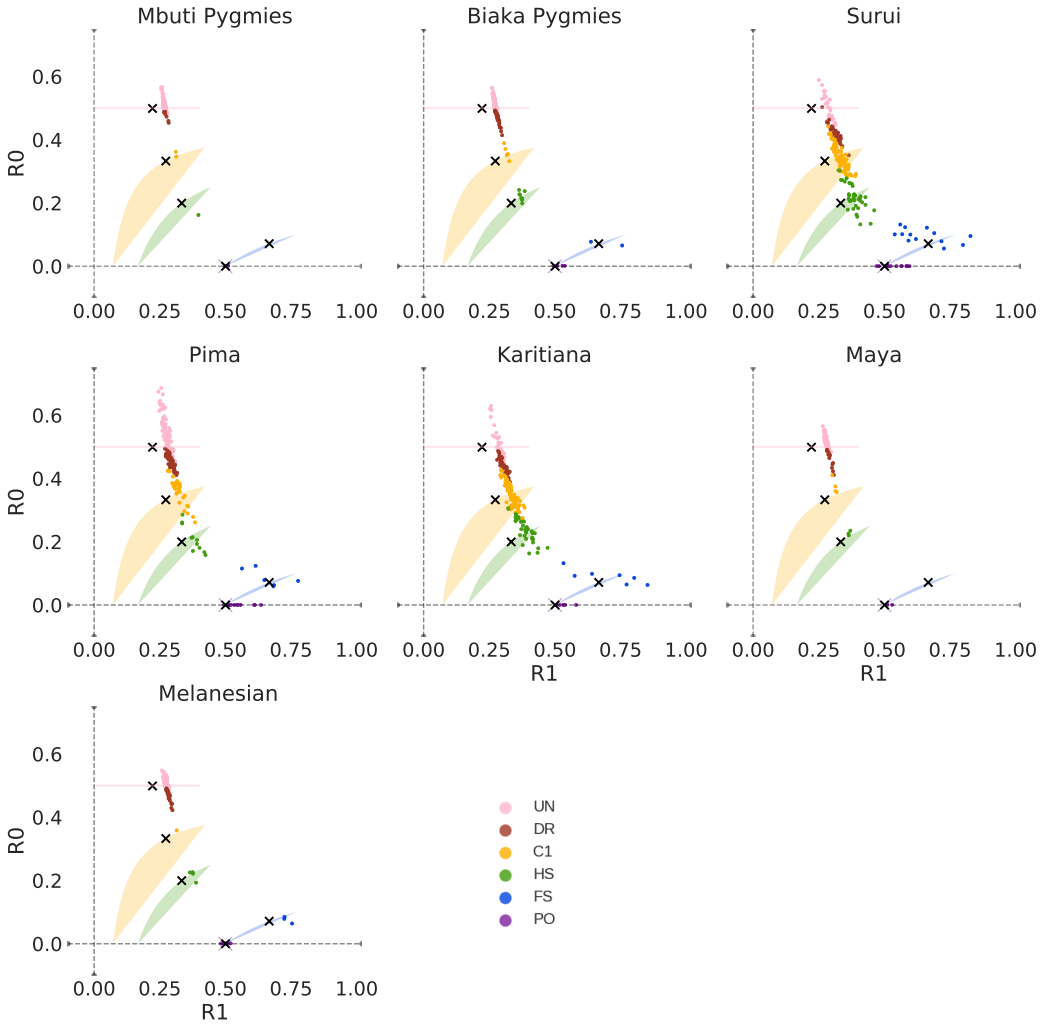


Figure S1: Scatterplot of R0 and R1 per HGDP population. Each pair of individuals within each population is represented by a point, which is colored according to the relationship category inferred using an allele frequency-based approach. The coloured shaded areas (sometimes just lines) show theoretically derived ranges of the joint expectation for specific relationship categories, as in Figures 2 and 3 in the main text. Black 'X's show values for pairs of individuals simulated under a constant N_e for each relationship category, as in Figure 2 of the main text. PO = parent-offspring, FS = full sibling, HS = half sibling, C1 = first cousin, DR = unknown relationship / distantly related, UN = unrelated.

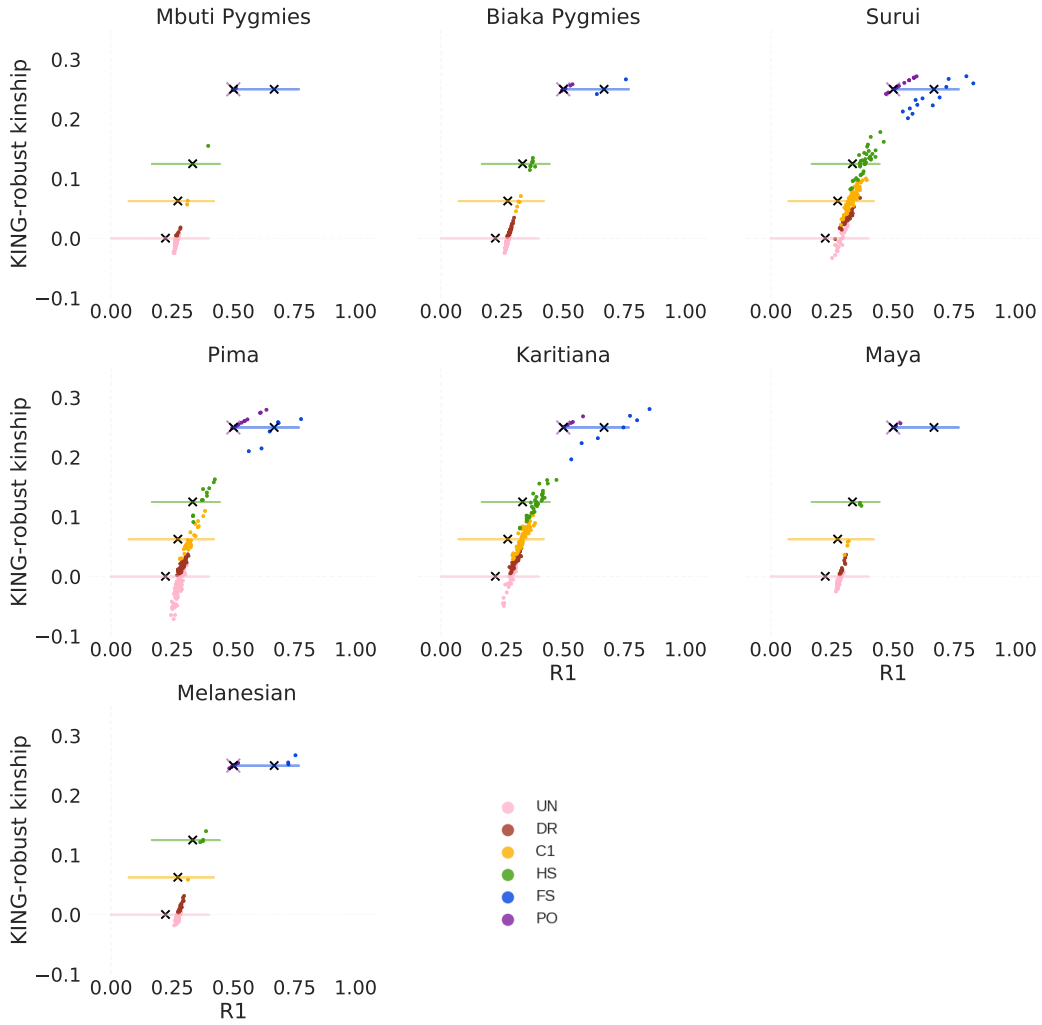


Figure S2: Scatterplot of R1 and KING-robust kinship per HGDP population. Each pair of individuals within each population is represented by a point, which is colored according to the relationship category inferred using an allele frequency-based approach. The coloured shaded areas (sometimes just lines) show theoretically derived ranges of the joint expectation for specific relationship categories, as in Figures 2 and 3 in the main text. Black 'X's show values for pairs of individuals simulated under a constant N_e for each relationship category, as in Figure 2 of the main text. PO = parent-offspring, FS = full sibling, HS = half sibling, C1 = first cousin, DR = unknown relationship / distantly related, UN = unrelated.

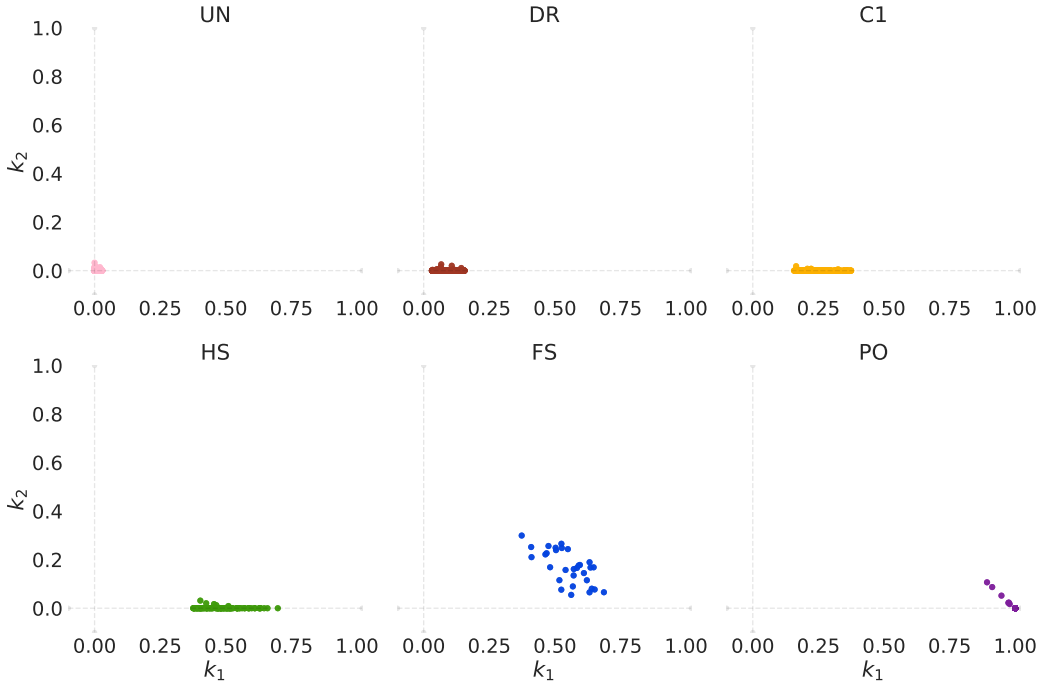


Figure S3: Scatterplot of the two relatedness coefficients k_1 and k_2 (denoted Z1 and Z2 in the output of PLINK) for each relationship category in the HGDP data. Estimates of the two relatedness coefficients are from the allele frequency-based approach implemented in PLINK. Each pair of individuals within each population is represented by a point, here they are paneled by the inferred relationship category: PO = parent-offspring, FS = full sibling, HS = half sibling, C1 = first cousin, DR = unknown relationship / distantly related, UN = unrelated. Also see figure 3 in the main text.

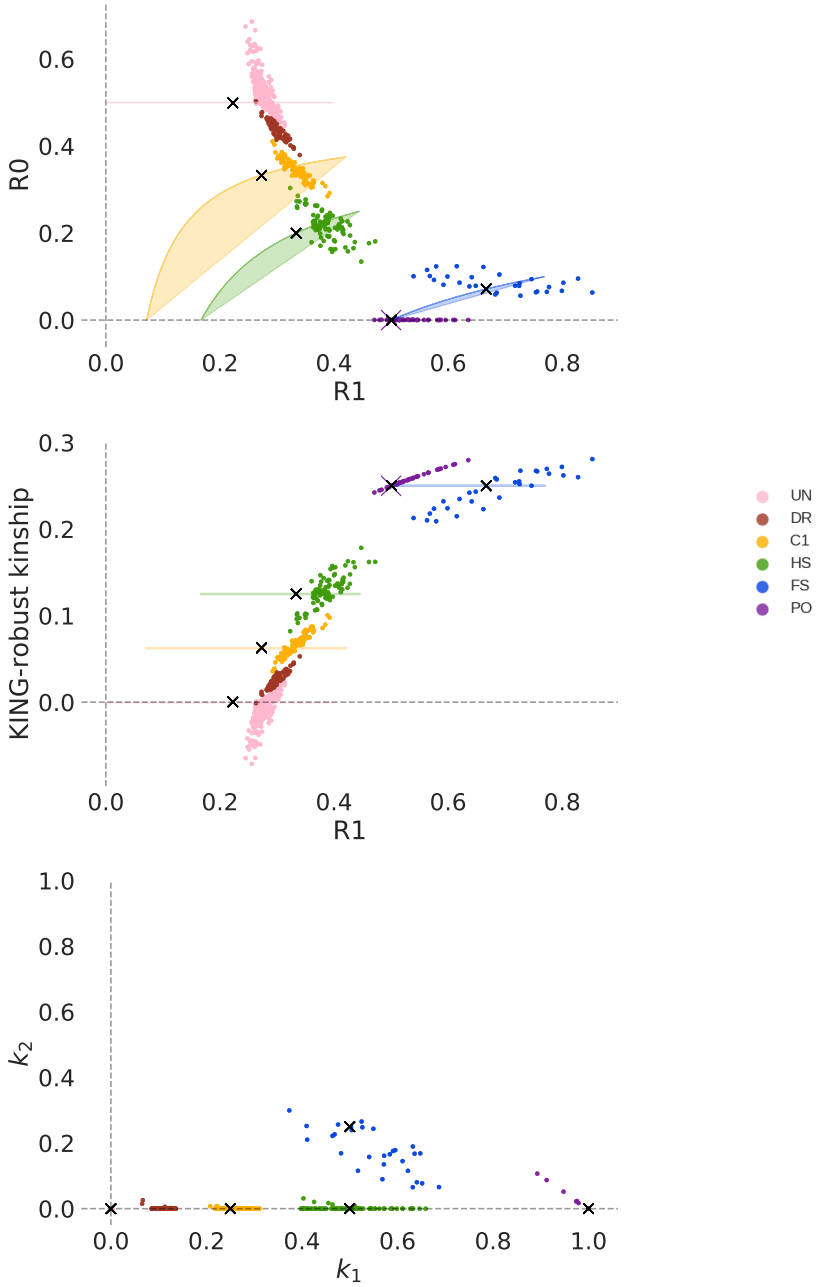


Figure S4: Pairwise scatterplots of R1, R0, and KING-robust kinship, and also k_1 vs k_2 when difficult to call relationships are excluded. Each pair of individuals within each population is represented by a point, which is colored according to the relationship category inferred using an allele frequency-based approach. Shaded areas and lines show the expected ranges for specific relationship categories, as in Figures 2 and 3 in the main text. PO = parent-offspring, FS = full sibling, HS = half sibling, C1 = first cousin, DR = unknown relationship / distantly related, UN = unrelated. Relationships were deemed difficult to call when the PLINK kinship was within 0.02 of the cutoff between two relationship categories.

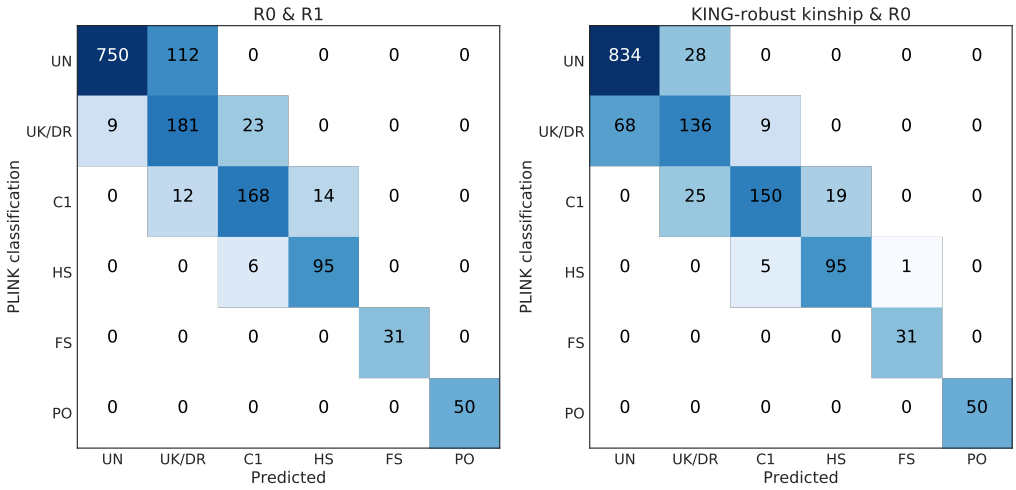


Figure S5: Confusion matrices for two classification schemes applied to the HGDP data. These matrices show concordance with the PLINK-based classification scheme described in the main text. Left: [R0, R1] Euclidean distance to data simulated under a constant N_e. Right: KING-robust kinship using the kinship criteria from Manichaikul et al. (2010), plus using R0 to distinguish PO from FS.

Tables below show the relationship-specific precision and recall for each classification.

	precision	recall	support
UN	0.99	0.87	862
UK/DR	0.59	0.85	213
C1	0.85	0.87	194
HS	0.87	0.94	101
FS	1.00	1.00	31
PO	1.00	1.00	50
avg/total	0.90	0.88	1451

	precision	recall	support
UN	0.92	0.97	862
UK/DR	0.72	0.64	213
C1	0.91	0.77	194
HS	0.83	0.94	101
FS	0.97	1.00	31
PO	1.00	1.00	50
avg/total	0.89	0.89	1451

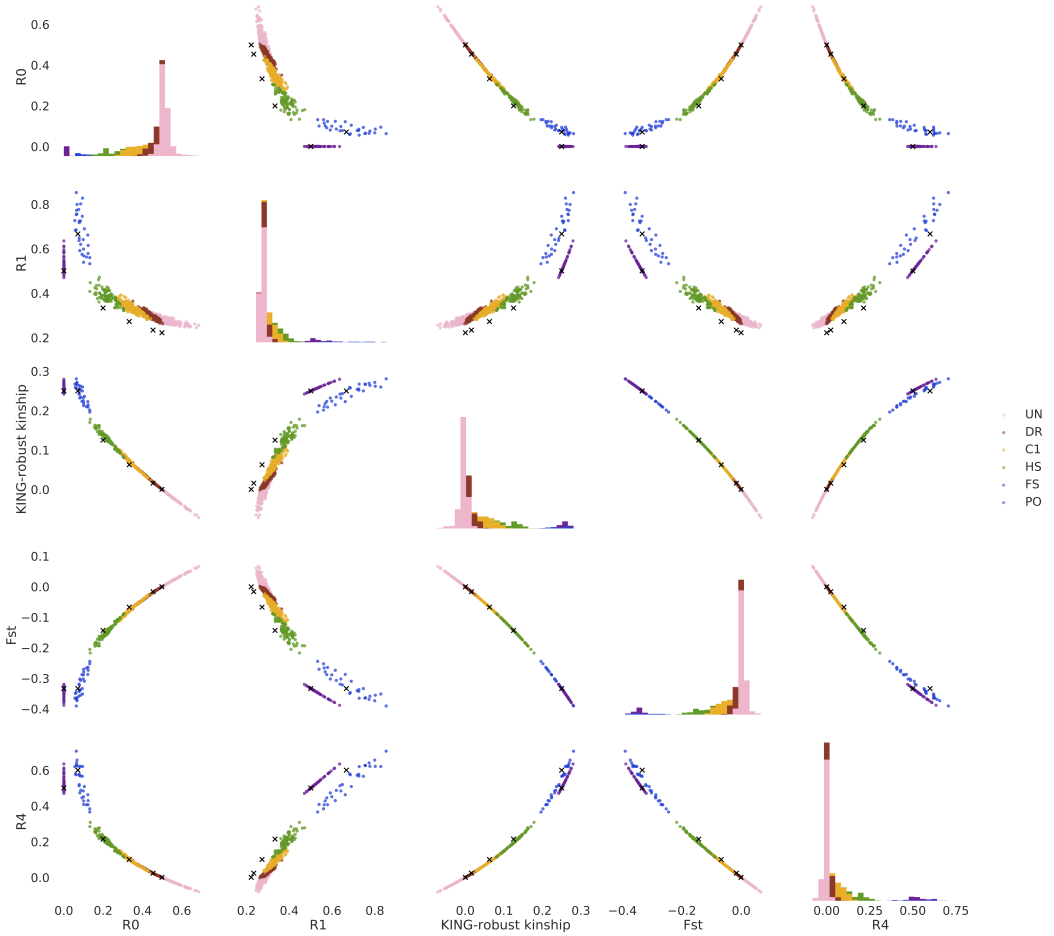


Figure S6: Scatterplots of R0, R1, KING-robust kinship, F_{ST}, and R4 for pairs of individuals within the selected HGDP populations. Each pair of individuals within each population is represented by a point, which is colored according to the relationship category inferred using an allele frequency-based approach. The coloured shaded areas (sometimes just lines) show theoretically derived ranges of the joint expectation for specific relationship categories, as in Figures 2 and 3 in the main text. Histograms of each statistic are on the diagonal. It is evident that pairs of statistics reduce the overlap in expected ranges between relationship categories. Also presented here are two related ratios not discussed in the main text: $F_{ST} = \frac{2C+2G-E}{2C+2G+B+D+E+F+H}$ and R4 $= \frac{E-2C-2G}{2C+2G+B+D+F+H}$. Black 'X's show values for pairs of individuals simulated under a constant N_e for each relationship category, as in Figure 2 of the main text. PO = parent-offspring, FS = full sibling, HS = half sibling, C1 = first cousin, DR = unknown relationship / distantly related, UN = unrelated.

2 Supplemental Texts

2.1 Text S1: Derivations of the expectations of R0, R1, and KING-robust kinship

We will here derive expressions for R0, R1 and KING-robust kinship for a range of different pairwise familial relationships. Based on these expressions we will then determine the joint range of expected values for R0 and R1 as well as R0 and KING robust kinship shown in figure 2 in the main text.

2.1.1 Assumptions and notation

In the below derivations will assume that we are analyzing data from two individuals, 1 and 2 that are not inbred and that are from the same homogeneous population. Additionally, we will assume that we have genotype data for n sites from both individuals and that all sites n sites are in Hardy-Weinberg equilibrium. In terms of notation, we will denote the two individuals' genotypes at given site s as g_1 and g_2, with g_1,$g_2 \in \{0, 1, 2\}$ corresponding to the number of copies of a specified allele (e.g., the derived allele) carried by individual 1 and 2, respectively. Also, we will denote the population frequency of the specified allele at site s as f. Furthermore, we will use the capital letters A through I to denote the probability of each of the nine different genotype pairs as shown in figure 1 in the main text. So e.g., A denotes the probability that both individuals have the genotype 0 at a site. Finally, we will use k_0, k_1, and k_2 to denote the probability that the two individuals share 0, 1 or 2 alleles identical-by-descent (IBD), respectively. The expected values of $K = (k_0, k_1, k_2)$ for different familial relationship can be seen in table S1.

Table S1: Expected $K = (k_0, k_1, k_2)$ for different relationship categories.

Relationship	k_0	k_1	k_2
Monozygotic twins (MZ)	0	0	1
Parent-offspring (PO)	0	1	0
Full siblings (FS)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Half siblings/avuncular/grandparent-grandchild (HS)	$\frac{1}{2}$	$\frac{1}{2}$	0
First cousins (C1)	$\frac{3}{4}$	$\frac{1}{4}$	0
Second cousins (C2)	$\frac{15}{16}$	$\frac{1}{16}$	0
Unrelated (UR)	1	0	0

2.1.2 Derivations of A through I

To derive the expected value of R0, R1 and KING-robust kinship for different familial relationship pairs, we first derive expressions for A to I, see also Toro et al. (2011) for a similar set of derivations, but notice they have a slightly different definition of k_1. We note that in general it must hold that in site s:

$$\begin{aligned} P(g_1, g_2|K = (k_0, k_1, k_2), f) &= P(g_1|f)P(g_2|g_1, K = (k_0, k_1, k_2), f) \\ &= P(g_1|f) \sum_{z\in\{0,1,2\}} P(Z = z|K = (k_0, k_1, k_2))P(g_2|g_1, Z = z, f) \\ &= P(g_1|f) \sum_{z\in\{0,1,2\}} k_z P(g_2|g_1, Z = z, f) \end{aligned}$$

where Z is an indicator of whether the two individuals share 0, 1 or 2 alleles IBD in a given site. Since we are assuming that the two individuals 1 and 2 are not inbred and that all sites are in Hardy-Weinberg equilibrium, the values of $P(g_1|f)$, $P(g_2|g_1, Z = 0, f)$, $P(g_2|g_1, Z = 1, f)$ and $P(g_2|g_1, Z = 2, f)$ must be those given in tables S2 to S5.

Based on this, we can derive expressions for A through I for an arbitrary degree of relatedness specified by K:

$$\begin{aligned}
A &= P(g_1 = 0, g_2 = 0|K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 0|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 0|g_1 = 0, Z = z, f) \\
&= f^2(k_0 f^2 + k_1 f + k_2 1) \\
&= k_0 f^4 + k_1 f^3 + k_2 f^2
\end{aligned}$$

$$\begin{aligned}
B &= P(g_1 = 1, g_2 = 0|K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 1|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 0|g_1 = 1, Z = z, f) \\
&= 2f(1-f)(k_0 f^2 + k_1 \frac{f}{2} + k_2 0) \\
&= k_0 2f^3(1-f) + k_1 f^2(1-f)
\end{aligned}$$

$$\begin{aligned}
C &= P(g_1 = 2, g_2 = 0|K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 2|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 0|g_1 = 2, Z = z, f) \\
&= (1-f)^2(k_0 f^2 + k_1 0 + k_2 0) \\
&= k_0 f^2(1-f)^2
\end{aligned}$$

$$\begin{aligned}
D &= P(g_1 = 0, g_2 = 1|K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 0|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 1|g_1 = 0, Z = z, f) \\
&= f^2(k_0 2f(1-f) + k_1(1-f) + k_2 0) \\
&= k_0 2f^3(1-f) + k_1 f^2(1-f) \\
&= B
\end{aligned}$$

$$\begin{aligned}
E &= P(g_1 = 1, g_2 = 1|K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 1|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 1|g_1 = 1, Z = z, f) \\
&= 2f(1-f)(k_0 2f(1-f) + k_1 \frac{1}{2} + k_2 1) \\
&= k_0 4f^2(1-f)^2 + k_1 f(1-f) + k_2 2f(1-f)
\end{aligned}$$

$$\begin{aligned}
F &= P(g_1 = 2, g_2 = 1|K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 2|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 1|g_1 = 2, Z = z, f) \\
&= (1-f)^2(k_0 2f(1-f) + k_1 f + k_2 0) \\
&= k_0 2f(1-f)^3 + k_1 f(1-f)^2
\end{aligned}$$

$$
\begin{aligned}
G &= P(g_1 = 0, g_2 = 2 | K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 0|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 2 | g_1 = 0, Z = z, f) \\
&= f^2(k_0(1-f)^2 + k_1 0 + k_2 0) \\
&= k_0 f^2 (1-f)^2 \\
&= C
\end{aligned}
$$

$$
\begin{aligned}
H &= P(g_1 = 1, g_2 = 2 | K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 1|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 2 | g_1 = 1, Z = z, f) \\
&= 2f(1-f)(k_0(1-f)^2 + k_1 \frac{(1-f)}{2} + k_2 0) \\
&= k_0 2f(1-f)^3 + k_1 f(1-f)^2 \\
&= F
\end{aligned}
$$

$$
\begin{aligned}
I &= P(g_1 = 2, g_2 = 2 | K = (k_0, k_1, k_2), f) \\
&= P(g_1 = 2|f) \sum_{z \in \{0,1,2\}} k_z P(g_2 = 2 | g_1 = 2, Z = z, f) \\
&= (1-f)^2(k_0(1-f)^2 + k_1(1-f) + k_2 1) \\
&= k_0(1-f)^4 + k_1(1-f)^3 + k_2(1-f)^2
\end{aligned}
$$

Table S2: $P(g_1|f)$.

$P(g_1 = 0\|f)$	$P(g_1 = 1\|f)$	$P(g_1 = 2\|f)$
f^2	$2f(1-f)$	$(1-f)^2$

Table S3: $P(g_2|g_1, Z = 0, f)$.

	$g_2 = 0$	$g_2 = 1$	$g_2 = 2$
$g_1 = 0$	f^2	$2f(1-f)$	$(1-f)^2$
$g_1 = 1$	f^2	$2f(1-f)$	$(1-f)^2$
$g_1 = 2$	f^2	$2f(1-f)$	$(1-f)^2$

Table S4: $P(g_2|g_1, Z = 1, f)$.

	$g_2 = 0$	$g_2 = 1$	$g_2 = 2$
$g_1 = 0$	f	$(1-f)$	0
$g_1 = 1$	$\frac{f}{2}$	$\frac{1}{2}$	$\frac{1-f}{2}$
$g_1 = 2$	0	f	$1-f$

Table S5: $P(g_2|g_1, Z = 2, f)$.

	$g_2 = 0$	$g_2 = 1$	$g_2 = 2$
$g_1 = 0$	1	0	0
$g_1 = 1$	0	1	0
$g_1 = 2$	0	0	1

2.1.3 Derivation of the expected values of R0

With the above expressions for A through I, we can derive an expectation of R0 for different relationships. We do this by first noting that:

$$\begin{aligned} R0 &= \frac{C+G}{E} \\ &= \frac{2C}{E} \\ &= \frac{2(k_0 f^2(1-f)^2)}{k_0 4f^2(1-f)^2 + k_1 f(1-f) + k_2 2f(1-f)} \\ &= \frac{2k_0 f(1-f)}{k_0 4f(1-f) + k_1 + 2k_2} \end{aligned}$$

Inserting the expected values of k_0, k_1 and k_2 from table S1 into this formula we get that if individuals 1 and 2 are a PO pair $R0$ is expected to be

$$\begin{aligned} R0_{PO} &= \frac{2 \times 0 f(1-f)}{0 \times 4f(1-f) + 1 + 2 \times 0} \\ &= \frac{0}{1} \\ &= 0 \end{aligned}$$

Similarly, for monozygotic twins (MZ), full siblings (FS), half siblings/avuncular/grandparent-grandchild (HS), first cousins (C1), second cousins (C2) and unrelated (UR) we expect $R0$ to be:

$$\begin{aligned} R0_{MZ} &= \frac{2 \times 0 f(1-f)}{0 \times 4f(1-f) + 0 + 2 \times 1} \\ &= 0 \end{aligned}$$

$$\begin{aligned} R0_{FS} &= \frac{2 \times \frac{1}{4} f(1-f)}{\frac{1}{4} \times 4f(1-f) + \frac{1}{2} + 2 \times \frac{1}{4}} \\ &= \frac{\frac{1}{2} f(1-f)}{f(1-f) + 1} \\ &= \frac{f(1-f)}{2f(1-f) + 2} \end{aligned}$$

which for $f \in]0, 1[$ has the range $]0, \frac{1}{10}]$.

$$\begin{aligned} R0_{HS} &= \frac{2 \times \frac{1}{2} f(1-f)}{\frac{1}{2} \times 4f(1-f) + \frac{1}{2} + 2 \times 0} \\ &= \frac{f(1-f)}{2f(1-f) + \frac{1}{2}} \end{aligned}$$

which for $f \in]0,1[$ has the range $]0, \frac{1}{4}]$.

$$\begin{aligned} R0_{C1} &= \frac{2 \times \frac{3}{4} f(1-f)}{\frac{3}{4} \times 4f(1-f) + \frac{1}{4} + 2 \times 0} \\ &= \frac{\frac{3}{2} f(1-f)}{3f(1-f) + \frac{1}{4}} \end{aligned}$$

which for $f \in]0,1[$ has the range $]0, \frac{3}{8}]$.

$$\begin{aligned} R0_{C2} &= \frac{2 \times \frac{15}{16} \times f(1-f)}{\frac{15}{16} \times 4f(1-f) + \frac{1}{16} + 2 \times 0} \\ &= \frac{30f(1-f)}{60f(1-f)+1} \end{aligned}$$

which for $f \in]0,1[$ has the range $]0, \frac{15}{32}]$.

$$\begin{aligned} R0_{UR} &= \frac{2 \times 1 \times f(1-f)}{1 \times 4f(1-f) + 0 + 2 \times 0} \\ &= \frac{2f(1-f)}{4f(1-f)} \\ &= \frac{1}{2} \end{aligned}$$

which is constant independent of the value for $f \in]0,1[$.

2.1.4 Derivation of the expected values of R1

With the above expressions for A through I we also get that

$$\begin{aligned} R1 &= \frac{E}{B+C+D+F+G+H} \\ &= \frac{E}{2B+2C+2F} \\ &= \frac{k_0 4f^2(1-f)^2 + k_1 f(1-f) + k_2 2f(1-f)}{2(k_0 2f^3(1-f) + k_1 f^2(1-f)) + 2(k_0 f^2(1-f)^2) + 2(k_0 2f(1-f)^3 + k_1 f(1-f)^2)} \\ &= \frac{k_0 4f^2(1-f)^2 + k_1 f(1-f) + k_2 2f(1-f)}{k_0(4f^3(1-f) + 2f^2(1-f)^2 + 4f(1-f)^3) + k_1(2f^2(1-f) + 2f(1-f)^2)} \\ &= \frac{k_0 4f(1-f) + k_1 + k_2 2}{k_0(4f^2 + 2f(1-f) + 4(1-f)^2) + k_1(2f + 2(1-f))} \\ &= \frac{k_0 4f(1-f) + k_1 + k_2 2}{k_0(4 - 6f(1-f)) + k_1 2} \end{aligned}$$

Inserting the expected values of k_0, k_1 and k_2 from table S1 into this formula we get that if individuals $i1$ and $i2$ are a PO pair $R1$ is expected to be

$$\begin{aligned} R1_{PO} &= \frac{0 \times 4f(1-f) + 1 + 0 \times 2}{0 \times (4 - 6f(1-f)) + 1 \times 2} \\ &= \frac{1}{2} \end{aligned}$$

Similarly, for MZ, FS, HS, C1, C2 and UR we expects $R1$ to be:

$$\begin{aligned} R1_{MZ} &= \frac{0 \times 4f(1-f) + 0 + 1 \times 2}{0 \times (4 - 6f(1-f)) + 0 \times 2} \\ &= \infty \end{aligned}$$

$$\begin{aligned} R1_{FS} &= \frac{\frac{1}{4} \times 4f(1-f) + \frac{1}{2} + \frac{1}{4} \times 2}{\frac{1}{4} \times (4 - 6f(1-f)) + \frac{1}{2} \times 2} \\ &= \frac{f(1-f) + 1}{2 - \frac{3}{2}f(1-f)} \end{aligned}$$

which for $f \in]0, 1[$ has the range $]\frac{1}{2}, \frac{10}{13}]$.

$$\begin{aligned} R1_{HS} &= \frac{\frac{1}{2} \times 4f(1-f) + \frac{1}{2} + 0 \times 2}{\frac{1}{2} \times (4 - 6f(1-f)) + \frac{1}{2} \times 2} \\ &= \frac{2f(1-f) + \frac{1}{2}}{3 - 3f(1-f)} \end{aligned}$$

which for $f \in]0, 1[$ has the range $]\frac{1}{6}, \frac{4}{9}]$.

$$\begin{aligned} R1_{C1} &= \frac{\frac{3}{4} \times 4f(1-f) + \frac{1}{4} + 0 \times 2}{\frac{3}{4} \times (4 - 6f(1-f)) + \frac{1}{4} \times 2} \\ &= \frac{3f(1-f) + \frac{1}{4}}{\frac{7}{2} - \frac{9}{2}f(1-f)} \end{aligned}$$

which for $f \in]0, 1[$ has the range $]\frac{1}{14}, \frac{8}{19}]$.

$$\begin{aligned} R1_{C2} &= \frac{\frac{15}{16} \times 4f(1-f) + \frac{1}{16} + 0 \times 2}{\frac{15}{16} \times (4 - 6f(1-f)) + \frac{1}{16} \times 2} \\ &= \frac{\frac{60}{16}f(1-f) + \frac{1}{16}}{\frac{62}{16} - \frac{90}{16}f(1-f)} \end{aligned}$$

which for $f \in]0, 1[$ has the range $]\frac{1}{62}, \frac{32}{79}]$

$$\begin{aligned} R1_{UR} &= \frac{1 \times 4f(1-f) + 0 + 0 \times 2}{1 \times (4 - 6f(1-f)) + 0 \times 2} \\ &= \frac{4f(1-f)}{4 - 6f(1-f)} \\ &= \frac{2f(1-f)}{2 - 3f(1-f)} \end{aligned}$$

which for $f \in]0, 1[$ ranges from $]0, \frac{2}{5}]$.

2.1.5 Derivation of the expected values of the KING-robust kinship estimator

Using the above expressions for A through I, the KING-robust kinship estimator (Manichaikul et al., 2010) can be re-written as:

$$
\begin{aligned}
KING\text{-}robust\ kinship &= \frac{E-2(C+G)}{B+D+H+F+2E} \\
&= \frac{E-4C}{2(B+F+E)} \\
&= \frac{k_1 f(1-f)+k_2 2f(1-f)}{2(k_0 2f(1-f)+k_1 2f(1-f)+k_2 2f(1-f))} \\
&= \frac{k_1 f(1-f)+k_2 2f(1-f)}{4f(1-f)(k_0+k_1+k_2)} \\
&= \frac{k_1 f(1-f)+k_2 2f(1-f)}{4f(1-f)} \\
&= \frac{k_1}{4}+\frac{k_2}{2}
\end{aligned}
$$

Hence, as expected, the expectation of the KING-robust kinship estimator is $\frac{k_1}{4}+\frac{k_2}{2}$ (which is the definition of kinship) regardless of the allele frequencies. Thus using the values in table S1 this means that the expected KING-robust kinship estimate is $\frac{1}{2}$ for MZ, $\frac{1}{4}$ for both PO pairs and full siblings, $\frac{1}{8}$ for HS, $\frac{1}{16}$ for C1, $\frac{1}{64}$ for C2 and 0 for unrelated pairs.

2.1.6 Joint ranges of R1 and R0

Above we derived the ranges of the expectation of each of R1 and R0 for different relationships. To get the joint ranges of the two, we note that the two ratios are not independent, because E is a part of both ratios. More specifically, (R1,R0) as a function of $f \in]0,1[$ for each of the different relationships considered here is shown by the solid lines in figure 2A in the main text. As this figure reveals, these are either single points (for PO) or concave, which means that for a combination of frequencies - and thus when more sites than one is considered - the ranges will be in the colored ranges inside the solid lines. It is important to note that these are simply ranges of expectations, because they are based on expected values of k_0, k_1 and k_2 for the different relationship. Hence, the realized values for a given pair will not necessarily lie inside the ranges shown as the realized values of k_0, k_1 and k_2 may differ from the expected values because the realized values for any related pair – except for parent off-spring and monozygotic twins – will vary around the expected values of k_0, k_1, and k_2 due to the randomness in the recombination process. E.g. a pair of half siblings are expected to have (k_0, k_1, k_2)=(0.5, 0.5, 0) but can in practice end up with e.g. (k_0, k_1, k_2)=(0.55, 0.45, 0) or (k_0, k_1, k_2)=(0.45, 0.55, 0). This will lead to values outside the expected range. In other words, the expectations derived here are expected values for our statistics in the same way as the values in table S1 are the expected values for K=(k_0, k_1, k_2).

2.1.7 Joint ranges of R1 and the KING-robust kinship estimator

Above we also derived the ranges of the expectation of each R1 and KING-robust for different relationships. We get the joint ranges from figure 2B in the main text by simply combining these. Again, it is important to note that these are simply ranges of expectations, because they are based on expected values of k_0, k_1 and k_2 for the different relationship. Hence, the realized values for a given pair will not necessarily lie inside the ranges shown as the realized values of k_0, k_1 and k_2 may differ from the expected values.

2.2 Text S2: The IBS method

In this section, we will describe technical details of the IBS method introduced in the main text. Specifically, this method aims to infer the frequency of different genotype combinations, p, for a pair of individuals, 1 and 2 from directly from sequencing read data. This is done using genotype likelihoods, i.e. probabilities of the read data given different genotypes, for each of the two individuals. These genotype likelihoods better reflect the uncertainty of the true genotypes that is inherent to low depth sequencing data. Briefly, this is accomplished by summing over all possible genotypes of the two individuals and weighting the probabilities using the corresponding genotype likelihoods. To fully describe the IBS method, we introduce the following notation:

- S denotes the number of sites
- $X_i = (X_i^1, X_i^2, ..., X_i^S)$ denotes the sequencing read data for individual i at the S sites
- $\mathcal{G} = \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$ denotes the set of possible genotypes
- Q_i^s denotes the (unknown) genotype of individual i at site s with $Q_i^s \in \mathcal{G}$
- $P(X_i^s|Q_i^s = q_i)$ is the likelihood of the genotype q_i for individual i at site s
- p is the vector of the frequencies of the genotype combinations that we aim to estimate

With this notation it must hold that:

$$\begin{aligned} P(X_1, X_2|p) &= \prod_{s=1}^{S} P(X_1^s, X_2^s|p) \\ &= \prod_{s=1}^{S} \sum_{(q_1,q_2)\in\mathcal{G}\times\mathcal{G}} P(X_1^s, X_2^s|(Q_1^s, Q_2^s) = (q_1, q_2)) \times P((Q_1^s, Q_2^s) = (q_1, q_2)|p) \\ &= \prod_{s=1}^{S} \sum_{(q_1,q_2)\in\mathcal{G}\times\mathcal{G}} P(X_1^s|Q_1^s = q_1) \times P(X_2^s|Q_2^s = q_2) \times P((Q_1^s, Q_2^s) = (q_1, q_2)|p) \end{aligned}$$

where the genotype likelihoods for the two individuals can be calculated using tools like ANGSD (Korneliussen et al., 2014) and where the probability of each possible genotype combination, $P((Q_1^s, Q_2^s) = (q_1, q_2)|p)$, is simply the element of p that corresponds to this genotype combination. This equation provides us with a likelihood function for the parameter, p, and we use this likelihood function to perform maximum likelihood estimation of p. In practice, this is done using an EM-algorithm which we have added to the software tool ANGSD with the name IBS.

We note that we have also added other similar models to ANGSD. The only difference between these and the model presented here is that fewer parameters are estimated, i.e. p is a shorter vector, and that $P((Q_1^s, Q_2^s) = (q_1, q_2)|p)$ is defined differently as a consequence. Those other models are not used in this paper.

2.3 Text S3: Supplemental Methods

2.3.1 Individuals excluded due to signs of admixture or inbreeding

We ran ADMIXTURE (Alexander et al., 2009) separately for each of the seven target populations. In each ADMIXTURE analysis we also include French and Han samples, to aid in identifying European or East Asian admixture, respectively. For the non-African target populations we also include Yoruban samples to identify African admixture. We excluded 16 samples with >5% contribution from more than once ancestry component. We estimated the inbreeding coefficient, F, for each of the remaining individuals using PLINK and excluded two individuals with f> 0.0625. This left us with a total of 142 individuals from the seven populations: Surui N=20, Pima = 20, Karitiana N=21, Maya N=16, Melanesian N=19, Biaka Pygmies N=31 and Mbuti Pygmies N=15.

2.3.2 Example command lines for IBS and SFS analyses

```
# make consensus - needed to make saf files
{ANGSD} -b ./data/1000G_aln/NA19042.mapped.ILLUMINA.bwa.LWK.low_coverage.20130415.
    list \
-r {CHR} -minMapQ 30 -minQ 20 -setMinDepth 3 -doFasta 2 -doCounts 1 -out ./data/
    consensus.NA19042.chr{CHR}

# make *.saf files (per individual)
{ANGSD} -b ./data/1000G_aln/NA19027.mapped.ILLUMINA.bwa.LWK.low_coverage.20130415.
    list \
-r {CHR} \
-ref ./data/1000G_aln/hs37d5.fa \
-anc ./data/consensus.NA19042.chr{CHR}.fa.gz  \
-sites ./data/1000G_aln/GEM_mappability1_75mer.angsd \
-minMapQ 30 -minQ 20 -GL 2 \
-doSaf 1 -doDepth 1 -doCounts 1 \
-out ./data/1000G_aln/saf/chromosomes/NA19027_chr{CHR}

# realSFS for each pair of individuals
{realSFS} ./data/1000G_aln/saf/chromosomes/NA19042_chr{CHR}.saf.idx ./data/1000
    G_aln/saf/chromosomes/NA19027_chr{CHR}.saf.idx -r {CHR} -P 2 -tole 1e-10 > ./
    data/1000G_aln/saf/chromosomes/NA19042_NA19027_chr{CHR}.2dsfs

# make genotype likelihood file
{ANGSD} -b ./data/1000G_aln/bamlist.all.txt \
-r {CHR} \
-sites ./data/1000G_aln/GEM_mappability1_75mer.angsd \
-minMapQ 30 -minQ 20 -GL 2 \
-doGlf 1 \
-out ./data/1000G_aln/GLF/chromosomes/chr{CHR}

# IBS
{IBS} -glf ./data/1000G_aln/GLF/chromosomes/chr{CHR}.glf.gz \
-seed {CHR} -maxSites 300000000 -model 0 \
-nInd 5 -allpairs 1 \
-outFileName ./data/1000G_aln/GLF/chromosomes/chr{CHR}.model0
```

2.3.3 Simulated ascertainment and demographic scenarios

Code conducting simulations, ascertainment, and analysis for Figure 2 is available at: `https://github.com/rwaples/freqfree_suppl`

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.

Toro, M. Á., García-Cortés, L. A., and Legarra, A. (2011). A note on the rationale for estimating genealogical coancestry from molecular markers. *Genetics Selection Evolution*, 43(1):27.

Paper II:
Where did the European ancestors of the Greenlanders come from?

By

Ryan K. Waples[1], Aviaja Lyberth Hauptmann[2,3], Inge Seiding[4]
...[0], Garrett Hellenthal[5], Torben Hansen[6]
Anders Albrechtsen[1,*], Ida Moltke[1,*]

0 Note that this author list is not complete. There will be a few more authors who contributed to data collection and basic study design, however they have not yet had a chance to read the manuscript presented here.

1 The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark

2 Ilisimatusarfik - The University of Greenland, Nuuk, Greenland

3 The Greenland Institute of Natural Resources, Nuuk, Greenland

4 Nunatta Katersugaasivia Allagaateqarfialu - Greenland National Museum and Archives, Nuuk, Greenland

5 UCL Genetics Institute (UGI), Department of Genetics, Evolution and Environment, UCL, London, UK

6 Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health, University of Copenhagen

∗ Corresponding Authors

Publication details

In preparation

WHERE DID THE EUROPEAN ANCESTORS OF THE GREENLANDERS COME FROM?

(IN PREPARATION)

Ryan K. Waples[1], Aviaja Lyberth Hauptmann[2,3], Inge Seiding[4], ...[0], Garrett Hellenthal[5], Torben Hansen[6], Anders Albrechtsen[1,9], and Ida Moltke[1,9]

[0] Note that this author list is not complete. There will be a few more authors who contributed to data collection and basic study design, however they have not yet had a chance to read the manuscript presented here.
[1] Section for Computational and RNA Biology, Department of Biology, University of Copenhagen
[2] Ilisimatusarfik - The University of Greenland, Nuuk, Greenland
[3] The Greenland Institute of Natural Resources, Nuuk, Greenland
[4] Nunatta Katersugaasivia Allagaateqarfialu - Greenland National Museum and Archives, Nuuk, Greenland
[5] UCL Genetics Institute (UGI), Department of Genetics, Evolution and Environment, UCL, London, UK
[6] Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health, University of Copenhagen
[9]Corresponding Authors

ABSTRACT

The Inuit ancestors of the Greenlandic people arrived in Greenland close to 1000 years ago. Since then, Europeans from many different countries have been present in Greenland during various periods of time, including Dutch whalers, Danish-Norwegian Lutheran and German Moravian missionaries, and Danish-Norwegian colonists. As a consequence, the current Greenlandic population has a substantial amount of genetic ancestry from Europe. In this study we investigate to what extent different European countries have contributed to the genetic makeup of the present-day Greenlandic people. From dense SNP chip data from 1582 admixed Greenlanders and a reference panel of 181 unadmixed Greenlanders and 8275 Europeans from 14 countries, we infer the ancestry sources of the European component of the Greenlandic genetic makeup. We use haplotype-based methods to obtain fine-scale resolution, which enables differentiation between European countries with genetically similar populations such as Denmark and Norway. Due to the rapid increase in population size in Greenland over the past 100 years we hypothesized that the earlier European interactions, such as the Dutch whalers and early missionaries, have contributed a larger amount of the European genetic ancestry compared to the recent Danish colonists. We find that the European ancestry within the Greenlandic people appears to be mainly Danish suggesting that most of the European admixture took place within the last few generations.

(WAPLES ET AL., IN PREPARATION)

Introduction

The Greenlanders are descendants of the Inuit of the Thule culture (Raghavan *et al.*, 2014) that entered Northern Greenland from Canada in the 12th century (Gulløv, 2008; Friesen and Arnold, 2008). When these Inuit ancestors arrived to Greenland the Norse had lived in the southern part of the island since 895 CE. The Norse stayed in Greenland until approx. 1450 CE, whereafter the Inuit were the only inhabitants of the island up until the arrival of British and Danish-Norwegian explorers starting in the 16th century (Gulløv, 2008). Neither archaeological nor genetic research show support direct contact or gene flow between the Norse and the Inuit population (Gulløv, 2008; Moltke *et al.*, 2015). However, from the 16th century many thousands of Europeans from various countries either visited or moved to Greenland. As a consequence there has been substantial gene flow from Europe into the Greenlandic population (Bosch *et al.*, 2003; Rasmussen *et al.*, 2010; Pereira *et al.*, 2015; Moltke *et al.*, 2015). A recent genetic study based on more than 10% of the adult population found that more than 80% of Greenlanders have some European ancestry and that Greenlanders have on average $\sim$25% European ancestry (Moltke *et al.*, 2015). Such genetic studies have given a detailed understanding of the Inuit ancestry of the present-day population of Greenland. However, there is a much less detailed understanding of the European component of the ancestry and it is not well known how much the different European countries have contributed to the Greenlandic gene pool.

European activities in Greenland after the Norse period ended can be roughly divided into four categories; scientific and trade expeditions, commercial whaling, missionary work, and colonization (see Figure 1). The search for the Northwest passage led English explorers to Canada and Greenland in the 1500's (Frandsen *et al.*, 2017). Subsequently, whaling became the main incentive for European activities in Greenland. For a long period in the beginning of the 17th century, whaling was predominantly led by England, The Netherlands, and Denmark-Norway, but eventually also Basque, French and German, especially Frisian, whalers were active around Greenland (Frandsen *et al.*, 2017; Brown, 1951). From 1670 onwards, the West Coast of Greenland became an important area for whaling and from around 1700, falling whale stocks in Svalbard made the Davis Strait attractive for European whalers, with whaling activities centered around Disko Bay, near present day Qeqertarsuaq and the shore towns of Ilulissat, Qasigiannguit and Aasiaat(Frandsen, 2010; Frandsen *et al.*, 2017). The Dutch were especially active whalers in Greenland, sending between 50-100 ships to Greenland each year from 1719 (Frandsen *et al.*, 2017).

Around the same time, in 1721, the arrival of the Danish-Norwegian missionary priest Hans Egede marked the beginning of the colonization of Greenland. This led to a new and more permanent type of contact between Inuit and Europeans, although whaling was still a primary draw with 107 Dutch ships and 14 German ships in West Greenland that year. In addition to the Danish-Norwegian missionaries, the German Moravian brethren established several missions in Greenland in the period 1733-1900, located in Nuuk and several other locations(Frandsen *et al.*, 2017). In contrast to other missionaries, most Moravian missionaries came in pairs as married couples and the mission had restrictions on intermarriage between the German missionaries and the Inuit members of their congregations (Wilhjelm, 2001).

In 1751 Denmark-Norway expanded colonial activities and claimed a monopoly on trade that sought to exclude other European nations from economic access to Greenland, particularly the Dutch (Frandsen *et al.*, 2017). Since then the primary contact between Greenlanders and Europeans has been with the countries within the Kingdom of Denmark. In particular, the Danish population, but populations of Iceland, previously a part of the Kingdom of Denmark, and the Faroe Islands, currently a part of the Kingdom of Denmark, have also historically interacted with Greenland and still today

(WAPLES ET AL., IN PREPARATION)

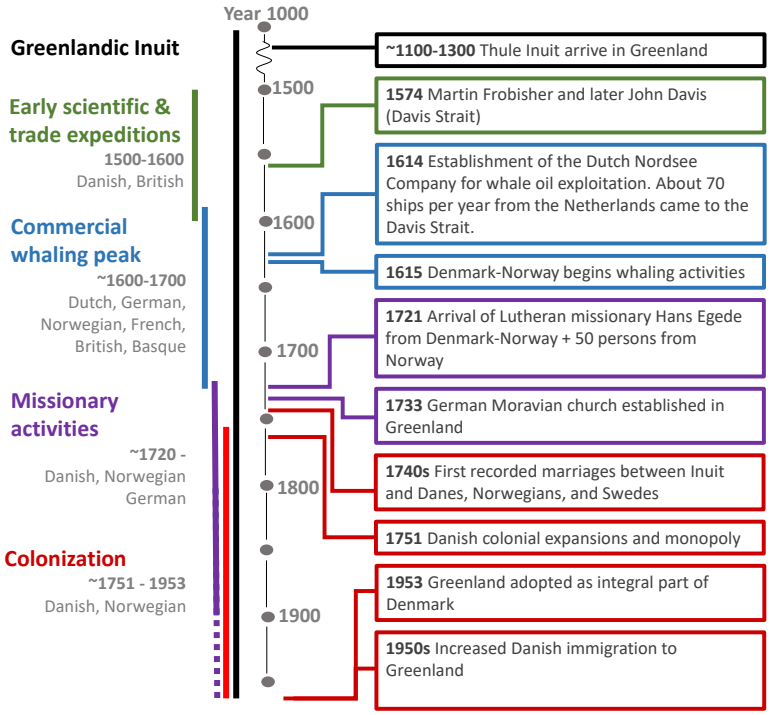


Figure 1: Timeline of selected significant events in Greenlandic-European contact.

make up a small fraction of immigration into Greenland (Statistics Greenland). Furthermore, it should be noted that Denmark-Norway was a conglomerate state until 1814, which means that people from the currently delimited groups 'Norwegian' and 'Danish' were both represented by the Kingdom of Denmark (Ostermann, 1940). Greenland remained a colony of Denmark until the 1950s, after which Greenland and its population has been considered an equal part of Denmark. The 1950's also marked the beginning of a period with a significant influx of Europeans. The migration consisted mainly of Danish workers but also, seasonally, of off-shore fishers primarily from Portugal and the Faroe Islands, using local harbours during the high seasons (Hansen, 1955, 1961; Frandsen *et al.*, 2017). Since 1979 Greenland has had its own parliament and in 2009 it established self-government, though it remains a part of the Kingdom of Denmark.

There is an extensive family registry from the Danish colonial period documenting personal relationships between the Inuit population of Greenland and European workers, traders and missionaries(Daveluy *et al.*, 2011; Seiding, 2011, 2016). The earliest marital contracts between Europeans and Greenlanders are from around the time of the establishment of the Danish trade monopoly in the early 1750s, but some marriages between Inuit and Danes, Swedes, and Norwegians in the 1740s are documented(Ostermann, 1940; Frandsen, 2010; Frandsen *et al.*, 2017). Church records, as well as census documents from the Royal Greenland Trading Department, contain detailed lists of the Inuit population as well as Greenlanders of both European and Inuit ancestry starting from the arrival of the missionaries and through much of the colonial period until the beginning of the 20th century (Seiding, 2016). At the beginning of the 19th century it was estimated that about 8% of the population had both Inuit and European ancestry (Seiding, 2013), but this estimate was based on only part of

(WAPLES ET AL., IN PREPARATION)

the population as a part of the archives including population counts were lost in the Hans Hedtoft shipwreck in 1959.

The degree of admixture prior to the colonial period is unknown and estimates are based on historical accounts and speculation. As an example, the whaling and trading activities of the Dutch has led to common beliefs that admixture with Dutch whalers was relatively common (Gad, 1969) resulting in a fraction of the population in whaling areas being of Dutch descent (Egedes, 2017). Notably, the Greenlandic population has grown dramatically from 10,000 at the beginning of the 1900s to more than 55,000 today (Hamilton and Rasmussen, 2010). Therefore, any European contribution to the gene pool before or at the beginning of the colonization would have had a much larger impact than more recent admixture.

Hence in summary the extent different European countries have contributed to the genetic makeup of the present-day Greenlandic people is an open question, which we aimed to address in this study. And prior to performing this study we hypothesised - based on the currently available knowledge - that not only Denmark but also The Netherlands, Norway and other countries who visited Greenland before the Danish colonisation has made a substantial contribution.

To address the question we have analysed genetic data from 3972 (1582 not closely related) Greenlanders, as well as data from 14 different European populations including Denmark, Norway, the Netherlands, and the UK. The detail to which it has been possible to address the European ancestry of the Greenlandic population has previously been limited due to the close genetic similarity between the European populations that have potentially contributed to the gene flow. However, using recently developed programs CHROMOPAINTER (Hellenthal *et al.*, 2014) and SOURCEFIND (Chacón-Duque *et al.*, 2018) we were able to disentangle the contribution from all of these highly genetically similar populations to the present-day Greenlandic population.

Materials and methods

Genotype datasets

Greenlandic data Study participants were Greenlandic individuals from two population surveys: the Inuit Health in Transition (IHIT) (n=3115) and a study consisting of Greenlanders living in Greenland (B99, n=1401), and Greenlanders living in Denmark (BBH, n=547) (Bjerregaard *et al.*, 2003; Bjerregaard, 2011). The cohorts have participants from 15 different locations in Greenland - from Qaanaaq in the northwest to Tasiilaq in the southeast (see Figure 2) as well as Greenlanders living in Denmark. All Greenlandic participants were genotyped on two SNP arrays: the CardioMetaboChip (196,224 SNPs) (Voight *et al.*, 2012; Moltke *et al.*, 2015) and the Multi-Ethnic Global Array (∼1.5M SNPs) citepBien2016-sh. Data from these two SNP arrays were merged on the plus strand and 3972 individuals with genotypes from both SNP arrays and a missing rate below 0.02 were retained. From these we removed singletons, sites not on an autosome, as well as sites with a significant (p<1e-10) deviation from Hardy-Weinberg equilibrium in a test that accounts for admixture (Meisner and Albrechtsen, 2019). The approval for population genetics analysis was given by the Commission for Scientific Research in Greenland (project 2014-08, 2014-098017).

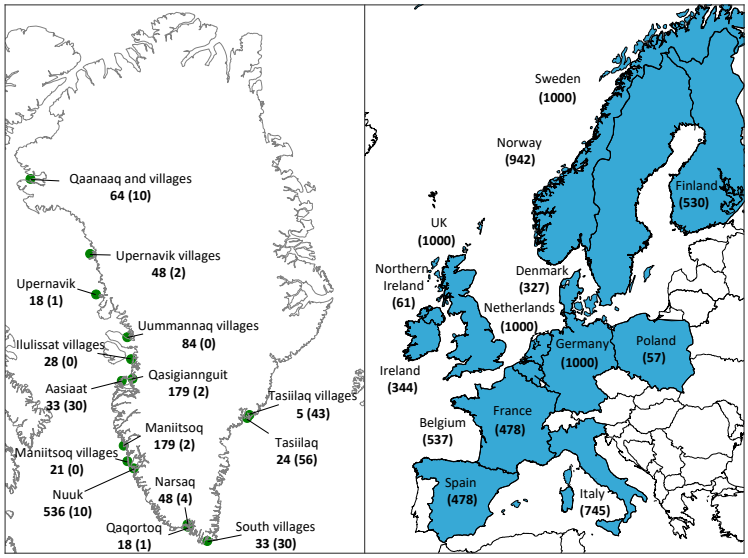


Figure 2: Overview of samples used in the CHROMOPAINTER analyses. Maps show sampling locations, sample sizes for each location is show in two parts, those in parentheses give the number of individuals included in the reference panel, outside parentheses give the number of admixed Greenlandic individuals included from each location in Greenland.

Greenlandic-European reference panel The Greenlandic genotype data (n = 3972) were merged with data from from 14 different European countries (n = 14,385): The UK, Sweden, Germany, Norway, Italy, Finland, Belgium, the Netherlands (Dutch), France, Ireland, Denmark, Spain, Northern Ireland, and Poland. The European data are from the Wellcome Trust Case Control Consortium (WTCCC2, EGAD00000000120, EGAD00010000124, EGAD00010000288, EGAD00010000632)

(Dubois *et al.*, 2010), and were selected to represent a broad spectrum of potential European admixture sources in Greenland (Figure 2).

European data sets were lifted to hg37 and put on the plus strand prior to merging. Also, we excluded sites within the MHC region and within the `HsInv0501` inversion on chr8, as well as sites with more than two alleles. Finally, we limited the number of individuals from each European country to 1000 and confirmed that there were no related individuals within each European country. When merging the Greenlandic and European datasets, we kept all sites present in both datasets and excluded 52 sites with more than 2% missing data. The resulting merged data set had 135,702 loci and 12,247 individuals with a total genotyping rate of 0.9995 and all loci with a minor allele count of at least 5.

1000 Genomes Data For admixture and local ancestry analyses (see below) we selected the Han Chinese in Beijing (CHB), Yoruba in Ibadan (YRI), and Utah residents with Northern and Western European Ancestry (CEU) population samples from the Thousand Genomes Project (1000G), for a total of 310 individuals. We used the phased genotypes from phase 3 aligned to GRCh37 from the vcf files available at http://www.internationalgenome.org.

Haplotype-based analyses The merged Greenlandic-European dataset was split by chromosome and phased without a reference panel using SHAPEIT (v2.r904) (Delaneau *et al.*, 2013) with default settings, using the HapMap phase II recombination map for hg37.

We excluded a number of Greenlandic individuals based on relatedness estimates and ADMIXTURE analyses (see separate subsections below). More specifically, we removed close relatives among all Greenlandic individuals by retaining at most one individual from each pair of individuals with a coefficient of relatedness > 0.2. Then we split the remaining Greenlanders into two sets based the results of a K=2 ADMIXTURE analysis: 1) the un-admixed Greenlanders with $> 99\%$ inferred Inuit ancestry, and 2) the admixed Greenlanders with $>1\%$ inferred European ancestry. From the second set we removed seventeen Greenlandic individuals estimated to have $>5\%$ African or $>7\%$ Asian ancestry in a K=4 ADMIXTURE analyses. These thresholds were selected to exclude individuals that differed markedly from the majority of other Greenlandic individuals (data not shown) and to be able to avoid having to include any Asian and African reference samples in our fine-scale analyses. We also excluded admixed Greenlandic individuals living in Denmark as these individuals may be more likely to have Danish ancestry than other European ancestries and we did not have similar samples of Greenlanders living abroad from other countries. This left us with a dataset consisting of 1582 not closely related Greenlanders with European admixture (target samples), 181 not closely related unadmixed Greenlanders (Inuit reference samples), and 8303 European reference samples.

Based on the results of a pilot CHROMOPAINTER analysis, we excluded 28 of the European reference samples because they were significant outliers (z-score > 5), based on comparing their chunkcounts to the rest of the individuals from their population (not shown). This resulted in a final set of 8275 European reference samples (Figure 2) and thus $8275 + 181 = 8456$ reference samples in total.

Analyses

ADMIXTURE We performed two different ADMIXTURE analyses to facilitate the generation of input data for all our main analyses: 1) an unsupervised K=2 ADMIXTURE analysis of all the 3972 Greenlandic individuals only assuming an Inuit and a European ancestry component, following a previous study (Moltke *et al.*, 2015) and 2) a supervised K=4 ADMIXTURE analysis of the Greenlandic individuals combined with individuals of European, Asian and African descent to investigate if there are any ancestry from Asian and African populations. In the K=4 analysis the Greenlandic individuals that were estimated to have >99% Inuit ancestry in the K=2 analysis served as references for the Inuit ancestry component.

Before the unsupervised K=2 analysis, we applied a minor allele frequency (MAF) filter of 0.05 to the Greenlandic data set described above (n = 3972), resulting in a dataset with 538,514 sites. For the supervised K=4 admixture analysis we merged the genotype data from the K=2 analysis with data from 310 individuals from three 1000G populations, retaining 521,622 overlapping sites after removing 46 sites with a greater than 0.25 frequency difference in the CEU compared to the European admixture component in the K=2 analysis. We selected the Han Chinese in Beijing (CHB), Yoruba in Ibadan (YRI), and Utah residents with Northern and Western European Ancestry (CEU) populations as proxies for Asian, African, and European ancestry, respectively.

For each of the two ADMIXTURE analyses we ran each ADMIXTURE v1.3.0 (Alexander *et al.*, 2009) ten times and selected the run with the maximum likelihood, checking convergence by ensuring multiple other runs within two log-likelihood units.

Relatedness To estimate relatedness coefficients for the Greenlandic samples we used relateAdmix (Moltke and Albrechtsen, 2014). This method accounts for admixture by estimating individual allele frequencies when estimating pairwise identity by descent (IBD) coefficients (k_1, k_2) based on genome-wide ancestry proportions for each individual. We used the K=2 genotype data and ADMIXTURE estimates of these genome-wide ancestry proportions. To estimate relatedness for the Europeans we applied the IBD inference function (-genome) in plink2 to the genotype data from all the Europeans.

CHROMOPAINTER We characterized the coancestry between Greenland and Europe with the haplotype-based method CHROMOPAINTER (Lawson *et al.*, 2012). This method is based on a Hidden Markov model (HMM) that statistically reconstructs ("paints") a target haplotype as a mixture of a set of reference haplotypes, allowing for recombination between the reference haplotypes. First, we painted each reference individual using all other reference individuals, then, we painted each admixed Greenlander using all reference individuals. We specified constant mismatch (Mu = 2.04 * 10e-5) and switch rate (N_e = 103.35) parameters across all analyses, which we estimated as the weighted mean values using data from chromosomes 1,4,15, and 22 in a subset of 168 individuals chosen to represent all reference populations, using 10 iterations of the expectation-maximization (EM) algorithm implemented in CHROMOPAINTER. For all CHROMOPAINTER analyses, we used the same recombination map as during haplotype phasing.

CHROMOPAINTER quantifies coancestry using two different measures, one based on the length of the genome copied from each donor in centiMorgans (cM), deemed "chunk lengths" by the program, and the second based on simple counting of the number of distinct ancestry chunks copied from each donor, deemed "chunk counts". Unless otherwise noted, we used the "chunk lengths" measure in downstream analyses. We summed the expected coancestry values across all chromosomes to produce a coancestry vector for each painted individual, giving the expected amount coancestry from

each reference individual. Notice, in this model, the coancestry between two individuals need not be symmetric.

Ancestry contributions from the European reference countries Based on the output from CHROMOPAINTER we estimated the ancestry contributions from each European reference country and Greenland with SOURCEFIND (v2) (Chacón-Duque *et al.*, 2018) and non-negative least squares regression (NNLS), as implemented in GLOBETROTTER (Hellenthal *et al.*, 2014). We applied both these methods in two ways: 1) to each admixed Greenlander individually 2) to the entire set of admixed Greenlanders as a group. The individual-based analysis allowed us to investigate the range of individual-level patterns of European ancestry, while the group analysis considers a large number of individuals at once, and estimates the ancestry sources of the mathematically-average admixed Greenlander.

To ensure convergence was reached in the SOURCEFIND analyses we ran 5 MCMC chains for each analysis and compared variance within and between separate chains with the Rhat diagnostic (Gelman and Rubin, 1992). Each of the chains were run with 1M iterations, a 100K burn-in and a thinning factor of 1000. We tested that we discarded enough to burn-in by computing Rhat while discarding the first 500K iterations, and compared these values to the shorter burnin (data not shown). For each ancestry source in each individual the Rhat diagnostic was consistent with MCMC convergence; mean Rhat across all chains was 1.0001, and the max value was 1.0044. For most of our SOURCEFIND analyses we used default priors with 8 eligible source and a mean of 4 sources expected to contribute. However, we tested if the results were robust to choice of prior by also running some additional analyses with a more sparse prior of 8 eligible source and a mean of 2 sources expected to contribute.

To assess the CHROMOPAINTER results we visualized the chunk counts coancestry matrix with PCA, following the fineStructure (Lawson *et al.*, 2012) documentation. We then summed the painting vector from each individual over the donor groups (e.g. countries). This reduced the length 8456 painting vector representing copying from individuals to a length 15 copying vector representing copying from each country. All PC analyses were conducted with sklearn.decomposition.PCA.

To assess the SOURCEFIND and NNLS results we evaluated our ability to identify ancestry associated with each reference group with a leave-one-out procedure. We inferred the ancestry of each reference individual while excluding them from the reference in the same way we analyzed each admixed Greenlander.

Investigating European admixture in the last few generations To investigate the timing of European admixture, we assigned local ancestry, either Inuit or European, in each admixed Greenlandic individual using RFMix (v2) (Maples *et al.*, 2013). In this analysis we used the same Inuit reference individuals as in the CHROMOPAINTER analysis, along with CEU individuals from 1000G to represent the European ancestry, this allowed us to utilize the larger number of overlapping loci with the 1000G data set. RFmix was run with default parameters, except we specified two different admixture dates, either 3 or 8 generations ago, to ensure that our results were robust to this choice. We used genotype data from the K=4 admixture analysis, with 521,622 sites, was split by chromosome and phased without a reference panel. After phasing, the reference Inuit and CEU individuals were used as the ancestry references for local ancestry inference in the admixed Greenlanders.

The length of admixture tracks is informative about the admixture date, but we found that the admixture was so recent, and the tracks were so long that the rate of phasing switch errors was large relative to the recombination rate since admixture, complicating the estimation of the ancestry tracks length distribution. Instead, we summarized the results for each individual by calculating the fraction

(WAPLES ET AL., IN PREPARATION)

of the genome, in cM, that has either two Inuit alleles, two European alleles, or one Inuit and one European allele. We found a few chromosomal regions, such as near the edge of chromosomes, with local ancestry fractions that were outliers relative to the rest of the genome, suggesting potential problems with the inference of local ancestry in these regions, or local genomic factors affecting ancestry. To address this, we removed 88 out of 26008 (0.3%) genomic windows of local ancestry calls with less than 62.5% Inuit ancestry or with more than 72.5% Inuit ancestry, for a total exclusion of 3.76 cM.

(WAPLES ET AL., IN PREPARATION)

Results

To investigate which European countries have contributed genetic ancestry to the current population of Greenland we analysed dense SNP array data from a large sample of not closely related, admixed Greenlanders (n = 1582) combined with a set of reference individuals aimed to represent the different possible ancestry sources for the admixed Greenlanders (Figure 2). The reference consisted of 181 unadmixed Greenlanders and 8275 individuals from 14 different European countries, including Denmark, Norway, Sweden, Germany, and the Netherlands (Figure 2). We first analysed this combined data set using haplotype-based methods to obtain estimates of the ancestry contribution from the 14 different European countries among Greenlanders. Then we performed several additional analyses to assess the validity of the results obtained. Finally, we looked into the timing of the admixture to further characterize the history of European admixture in Greenland.

Inference of European admixture sources with haplotype based analyses

We performed the haplotype-based analyses in two steps. The first step was to use CHROMOPAINTER, to reconstruct ("paint") the genomes of the admixed Greenlandic individuals as a mixture of the haplotypes in our reference samples, providing an estimate of their co-ancestry with these reference samples. The second step was to apply SOURCEFIND, a Bayesian MCMC-based method, to estimate the genetic contribution of the Greenland Inuit and each of the 14 different European countries to the ancestry of the admixed people in Greenland. We conducted this latter analysis in two ways: to each admixed Greenlandic individual independently, and to all the admixed Greenlanders as a group.

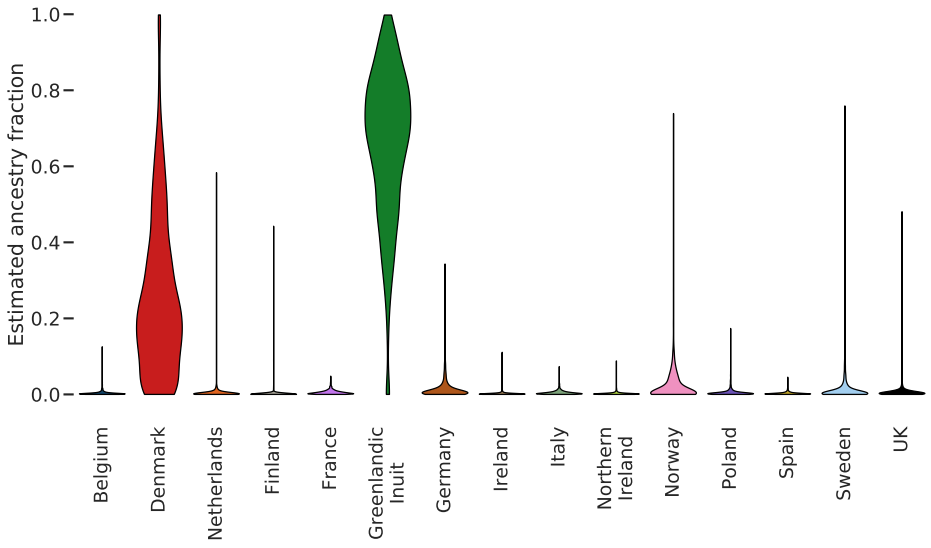


Figure 3: Violin plot of the per-country ancestry estimates by SOURCEFIND, based on the individual analysis. Each source country has a violin showing the distribution of the estimated mean ancestry fraction from that country, across all admixed individuals. Each admixed individual appears in the distribution for each country. The mean ancestry for each individual is calculated across all MCMC iterations.

(WAPLES ET AL., IN PREPARATION)

	Individual-based			Group-based	
	>=1%	>=5%	>=20%	SOURCEFIND	NNLS
Belgium	0.0% (0)	0.0% (0)	0.0% (0)	0.1%	-
Denmark	76.4% (1208)	69.5% (1100)	35.8% (567)	31.6%	27.9%
Dutch	0.1% (1)	0.1% (1)	0.1% (1)	0.1%	-
Finland	0.6% (9)	0.3% (5)	0.1% (1)	0.0%	-
France	0.0%	0.0% (0)	0.0% (0)	0.2%	0.6%
Greenlandic Inuit	98.3% (1555)	98.3% (1555)	97.4% (1541)	65.6%	64.9%
Germany	0.1% (1)	0.1% (1)	0.0% (0)	0.3%	0.8%
Ireland	0.2% (3)	0.2% (3)	0.0% (0)	0.1%	-
Italy	0.0% (0)	0.0% (0)	0.0% (0)	0.2%	1.3%
Northern Ireland	0.0% (0)	0.0% (0)	0.0% (0)	0.1%	-
Norway	17.8% (281)	6.2% (98)	1.1% (18)	0.7%	2.6%
Poland	0.1% (2)	0.1% (1)	0.0% (0)	0.2%	0.7%
Spain	0.0% (0)	0.0% (0)	0.0% (0)	0.2%	-
Sweden	3.5% (56)	1.3% (20)	0.1% (1)	0.3%	0.9%
UK	0.3% (5)	0.2% (3)	0.1% (1)	0.2%	0.2%

Table 1: Assignment to country at 1%, 5%m and 20% ancestry thresholds across 1582 admixed Greenlanders. The values shown are percentages (counts) of the number of individuals inferred to have at least [1%, 5%, 20%] ancestry, arranged by column from each source country. To be counted here, an individual must have had at least [1%, 5%, 20%] ancestry with a posterior probability above 99%, see supplement for a table including results based on posterior probability thresholds of 95% and 99.9%.

When performing inference on each admixed individual, we estimated the 1582 admixed Greenlanders to have an average of 65.8% Inuit ancestry, and 34.1% European ancestry (Figure 3). In total we found 1100 individuals (69.5%) assigned at least 5% Danish ancestry with 0.99 posterior probability, the most of any European country (Table 1). The other European countries found to contribute more than 5% ancestry to five or more individuals are all Nordic countries: Norway with 98 (6.2%), Sweden with 20 (1.2%), and Finland with 5 (0.3%). We find only very few individuals with high posterior probability (>.99) of having more than 5% ancestry from other regions of Europe including the British-Irish Isles (0.3%), and the Netherlands/Belgium (0.06 %). The same overall pattern is observed when a lower 1% ancestry threshold is used instead of 5% (Table 1) and if we use different posterior probability thresholds (Supplemental Table 2). With an ancestry threshold of greater than 20%, there is essentially only European ancestry from Denmark and Norway (Table 1). In locations within Greenland, European ancestry is less common in the North and East, as noted by previous studies (Moltke *et al.*, 2015), and visible here as an elevated number of reference Inuit individuals in Qaanaaq and Tasiilaq. Second, almost all of the few individuals with a posterior probability above 0.99 of having at least 5% ancestry from the UK or Ireland are from the Northern part of Greenland despite the fact that only 4% of the 1582 admixed Greenlandic individuals are from there (Figure 2).

When performing the group-based analysis, the estimated Inuit ancestry fraction is consistent (65.6%) and the Danish ancestry component is estimated to be 31%, with no other European country contributing more than 1%. In this analysis, Denmark makes up 91% of the total estimated European ancestry, with the only other country contributing more than 1% of the European ancestry being Norway at 2.1% (Table 1).

Taken together, the two analyses both suggest that the vast majority of the European ancestry among Greenlanders is Danish, with a smaller fraction of it inferred to be from other Nordic countries, especially Norway, and very little from the UK and the Netherlands.

(WAPLES ET AL., IN PREPARATION)

Validity of results

To investigate the validity of our results, we performed several additional analyses. First, we tried to assess to what extent the CHROMOPAINTER and SOURCEFIND methods make it possible to sensibly distinguish between ancestry from the different European countries based on the available data. For CHROMOPAINTER we did this by performing principal component (PC) analysis of CHROMOPAINTER co-ancestry estimates for all 10038 individuals in our dataset, after summing over the ancestry contributions from each country (Figure 4, Supplemental Figure 4). The first PC axis separates the reference Inuit from all the Europeans, with the admixed Greenlanders falling between the Europeans and the reference Inuit. Subsequent axes tend to separate the individuals from one or two European countries from the rest. For example, the third PC axis separates the Norwegian individuals from most of the other countries, including Denmark, and the ninth and tenth PC axes combined separate the Danish individuals, suggesting that the CHROMOPAINTER results allows us to tell apart individuals even from countries like Norway and Denmark that are genetically very similar. Notably, the many admixed Greenlanders have a pronounced weighting on the ninth and tenth PC axes, projecting them coincident with the Danish individuals. And along the third PC axis, which separates the Norwegian reference samples from Danish, most of the admixed Greenlanders fall on top of the Danish individuals. Hence, the CHROMOPAINTER output also seems to support the notion that the European ancestry of the Greenlanders to a large extent is from Denmark as opposed to Norway.

We assessed assignment to country using a leave-one-out strategy: we inferred the ancestry of each reference individual with SOURCEFIND and NNLS, while excluding them from the reference, checking to what extent the results matched the individuals's known country of origin. This demonstrated a relatively high ability to identify ancestry from most countries, especially Denmark (89%), the Netherlands (87%) and Norway (90%). Importantly, we did not observe high rates of misassignment between Denmark, Norway, and the Dutch, which were all below 3%, suggesting it is possible to distinguish these ancestry sources.

Second, we ensured the results were consistent with alternative methods. In particular, we performed an ADMIXTURE analysis and found that for all admixed Greenlanders, the Inuit ancestry proportion estimated using ADMIXTURE was highly correlated both with the percent of genome (in cM) that CHROMOPAINTER inferred to be copied from Inuit and with the individual-based Inuit ancestry fractions estimated by SOURCEFIND (Supplemental Figure 2). We also ran NNLS as implemented in GLOBETROTTER and found results that are qualitatively similar to the results of SOURCEFIND (Table 1). However, the leave-one-out analysis with NNLS showed a reduced ability to recover country of origin (Supplemental Figures 5 & 6). Finally, we reran SOURCEFIND with another choice of priors than the default ones used in our initial analyses. This also resulted in qualitatively very similar results (Supplemental Table 4).

Investigating European admixture in the last few generations

To further characterise the European admixture history in Greenland, we performed an analysis to investigate the timing of admixture in Greenland. Specifically, we applied RFmix to the set of admixed Greenlander to estimate the proportions of the genome where each individual of interest has 1) inherited both alleles from Inuit ancestors, 2) inherited both alleles from European ancestors or 3) inherited one allele from an Inuit ancestor and one allele from a European ancestor. Like the genome-wide proportion of Inuit ancestry these three "ternary ancestry fractions" should be robust to phasing switch errors because they do not rely on the length of ancestry tracts and importantly

(WAPLES ET AL., IN PREPARATION)

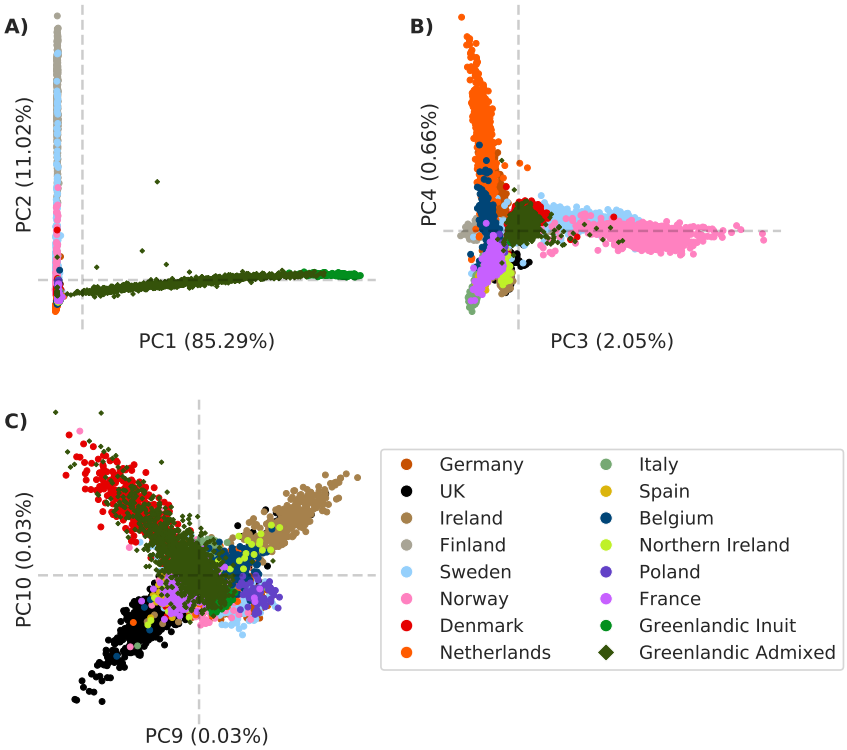


Figure 4: Scatter plots of selected principal component (PC) axes. Colored shapes represent reference (circle) and admixed (diamond) individuals, and are shaded by country. The percentages given on each axis show the percentages of variance explained. PC analysis conducted on the CHROMOPAINTER coancestry matrix, summed over the reference individuals from each country. Plotting z-order is the same as the legend. Further PCA plots are available in the supplement. **A)** PC1 - 2, **B)** PC3-4, PC9-10.

they contain information about the timing of admixture. For example, an absence of sites where both alleles are inherited from Inuit ancestors suggests that at least one parent is of European ancestry which in turn suggests admixture with a European took place in the last generation (blue axis on Figure 5A). And similarly other configurations are expected for different other recent admixture events like first, second, and third generation European admixture (see Figure 5A for details).

When we summarized the ternary ancestry fractions of the 1582 admixed Greenlanders (Figure 5B), we saw several interesting patterns. First, along the left edge, we find 250 admixed Greenlanders consistent with having at least one unadmixed European parent (blue and yellow dots in Figure 5B). Of these, 27 have two European alleles at nearly every position (yellow dots on Figure 5B), suggesting they have two European parents each. Together the 277 European parents of these 250 individuals account for $>$8% of the overall ancestry of the admixed individuals (277/(2*1582)). Assuming a European ancestry fraction to be $\sim$35%, this set of parents accounts for $\sim$25% (8/35) of the total European ancestry we observe. Notice, the 250 individuals are pulled slightly off the left axis, we believe this is due to some minor bias in our local ancestry inference. The alternative explanation would be European individuals with small amounts of Inuit admixture, occurring many generations ago. However, we do not find signs of numerous individuals with this ancestry pattern in any other analyses, so we found it more likely that it is caused by a slight bias in our local ancestry inference.

(WAPLES ET AL., IN PREPARATION)

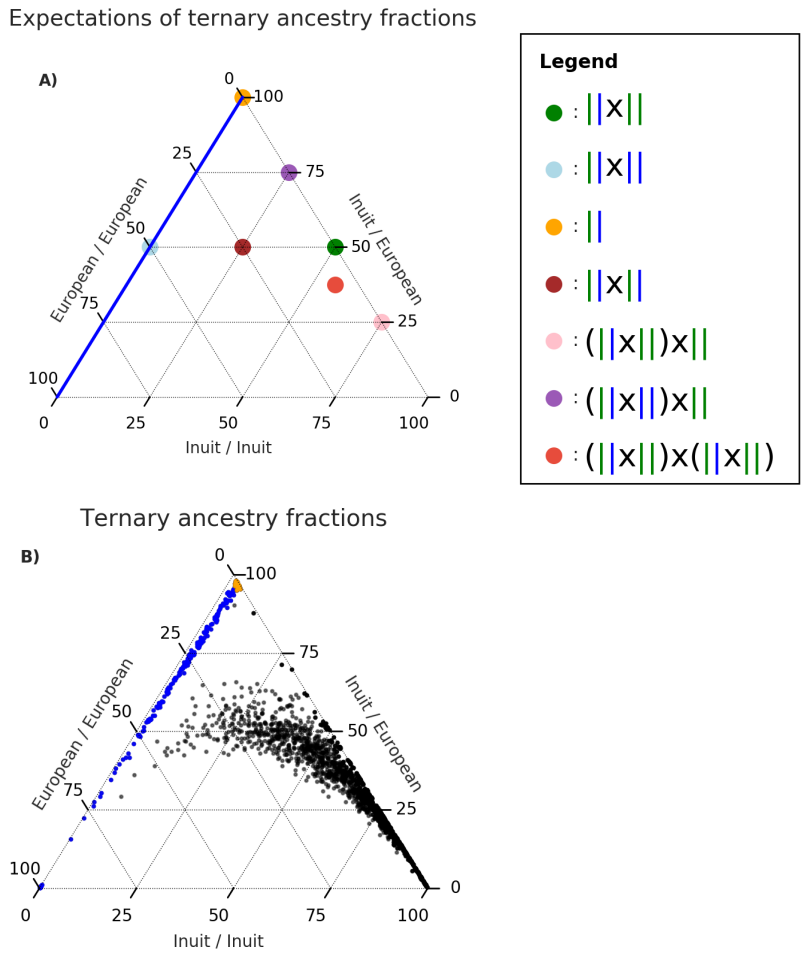


Figure 5: Ternary plots of "ternary ancestry fractions" i.e. the genome-wide fraction of the genome where 1) both alleles have Inuit ancestry (shown on the Inuit/Inuit axis), 2) both alleles have European ancestry (European/European axis) or 3) one allele has Inuit ancestry and one has European ancestry (European/Inuit axis). The relative abundance of different fractions is informative about admixture history. A) A ternary plot where the dots show the expectations for the three ternary ancestry fractions for a number of possible recent admixture histories between Greenlanders and Europeans. The three corners of the plot represent genomes with all loci having two European alleles (bottom left), two Inuit alleles (bottom right), or one Inuit and one European alleles (top). The left axis in blue indicates fractions that are expected for individuals with at least one fully European parent since it has very few sites with two Inuit alleles. The colored dots show expected ternary fractions for selected admixture histories: individuals with one unadmixed Greenlandic parents and one European parent (yellow), three Inuit grandparents and one European grandparent (green), three European grandparents and one Inuit grandparent (blue), one Inuit and one European grandparent on each side (brown), three Inuit and one European great-grandparent on one side, and an Inuit parent on the other (pink), and finally three Inuit and one European great-grandparent on each side (red). **B)** Ternary plot with a dot for each of the 1582 admixed Greenlanders showing their estimated ternary ancestry fractions. The colors convey the way we have categorized the individuals in the text.

(WAPLES ET AL., IN PREPARATION)

Second, many of the remaining admixed Greenlanders have ternary ancestry fractions close to the expected values from second and third generation admixture with Europeans shown in Figure 5A (dots near the right axis, Figure 5B). However, it is important to emphasize that a lot of these fractions could also be the result of older admixture, so this pattern has to be interpreted with caution.

A last pattern worth noticing is that there is a high degree of variation in ternary ancestry fractions, which is not expected from a history with only older admixture.

Taken together, the observed ternary fractions are consistent with an admixture history of mainly admixture within the last few generations, but some older gene flow as well. This history matches a time when Denmark, according to historical records, accounted for a large portion of the European contact, trade, and migration with Greenland. And in turn, this makes our finding of mainly Danish ancestry consistent with historical records.

(WAPLES ET AL., IN PREPARATION)

Discussion

The genetic ancestry of present-day Greenlanders is mainly Inuit, but also to a large extent European. Here we further investigated which European countries have contributed to this European ancestry by applying powerful haplotype-based methods to genetic data from almost 2000 not closely related Greenlandic individuals from 15 towns and villages across Greenland and more than 8000 European individuals from 14 countries. We found that the majority of the European ancestry among Greenlanders originates from Denmark from admixture events within the last few generations; a period of time when Danes, according to historical records, has accounted for a large portion of the European contact, trade, and migration with Greenland.

However, there are several issues worth reflecting on. In particular, a major consideration in genetic assignment analyses is the extent to which the samples included in the analyses are representative of the groups we want to draw inferences about. In our case, we have to consider 1) if the included Greenlandic individuals are representative of Greenland as a whole, 2) if the included European individuals are representative of their respective countries, 3) if the reference sample we used includes all potentially relevant ancestry sources and 4) whether countries are suitable ancestry groups for our analyses.

In this study we have analysed a large sample of the Greenlandic population selected as part of a public health survey to represent all of Greenland. However, we did not have participants from all inhabited places in Greenland. In particular, we lack individuals from Disko Island, the center of Dutch whaling activities in Greenland (Frandsen *et al.*, 2017). It is certainly possible we may have found more Dutch ancestry in individuals from Disko than we did in the individual we considered in this study. That said, while it would be very interesting to have data from the unsampled locations, the individuals considered here represent an extensive sampling from Greenland, and thus the results suggest that if the Dutch have indeed contributed substantially to the Greenlandic population they have only done so very locally and their descendants have remained in place.

Assessing whether the samples from the European countries are representative for those countries is difficult, because we we do not have access to the within-country metadata. However, many of the same European individuals have previously been used for a similar study of the fine-scale ancestry of the British Isles (Leslie *et al.*, 2015).

Regarding the related question of whether the reference we used includes all potentially relevant ancestry sources, we did include samples from almost all the European countries with historical record of contact since 1500 (Figure 1). There are only two exceptions. One of these is Portugal, which we were not able to acquire publicly available samples from. The historical data about contact with Portugese mainly suggest an influence from the seasonal offshore cod fisheries in the period between 1930-1970. As the colonial monopoly was gradually loosened after WW2, foreign fisheries in Greenlandic waters began and then peaked in the 50's and 60's with the Portuguese as the most active and successful nation in the fisheries (Hansen, 1955, 1961). However, contact to the foreign fishermen was limited to the occasional contact when the ships docked at Greenlandic harbours. However, we expect the lack of such samples to have a minimal effect because we included samples from Spain, and saw very little sign of Spanish ancestry.

The other exception is Iceland and the Faroe islands, that are island nations whose position in the North Atlantic facilitated contact with Greenland. They would be difficult to include in our analysis as potential source countries due to their recent history of admixture and gene flow from the source countries already included in this study (Ebenesersdóttir *et al.*, 2018). For this reason, it is difficult to

draw any conclusions on whether the European ancestors of any Greenlanders lived in one of these countries for a few generations prior to immigration to Greenland.

With respect to the fourth and last issue, we chose to use present-day countries as the the unit of analysis, in the sense that we describe each individual as having a mixture of ancestries from different countries. Previously published CHROMOPAINTER analyses have often not used present-day countries as units, but rather defined ancestry groups based on a genetic clustering of individuals. For example, Chacon-Duque et al. (2018) defined 56 "surrogate" ancestry sources from across the world and used them to describe admixture in Latin Americans. We chose to use countries for a two reasons. First, countries are easy to interpret and immediately recognizable. Second, the recent timescale of much of the admixture, suggest that current day countries are reasonable proxies for the ancestry at the time of admixture. The results from the leave-one-out analyses suggest that this choice was reasonable and that it is possible to distinguish between the reference countries. In particular, we demonstrated that we could tell apart individuals from key countries such as Denmark, Norway and the Netherlands and that the results were fairly robust to the methodology used.

Despite these shortcomings, the results consistently point towards the European ancestry of the present-day Greenlanders being predominantly Danish and recent. This is somewhat unexpected and indicates that European activities prior to Danish colonization did not have a significant impact on the genetic composition of the population in Greenland. This is in contrast both to common beliefs in Greenland (Gad, 1969) and our own initial hypothesis. The lack of admixture from whaling countries, especially the Netherlands is noteworthy, but might be explained by a number of factors. First, whaling activities occurred mostly near Disko Bay, contact with the local population was sporadic and whalers did not commonly spend the winter in Greenland, but instead returned to Europe. Second, Dutch and other European whaling activities in Greenland were limited to about a century after which the Danish colonization and economic monopoly ensured that Denmark was major point of contact between Europe and Greenland. Finally, it has been postulated that first contact with Europeans was followed by severe epidemics and that the interaction with the Dutch around Disko probably led to some of the first incidents of tuberculosis in the region (Gad, 1969). A well-documented example was a severe smallpox epidemic in Nuuk in the 1730s following the arrival of European ships (Gad, 1969). As most victims of the epidemics were women and children and particularly those with close contact with Europeans this would have led to fewer survivors among admixed children in the beginning of contact with Europeans as compared to later contact with Europeans, which was highly dominated by Danish. Among other Europeans with prolonged contact with the Greenlandic population were the German Moravian brethren, who stayed in Greenland for about 170 years until 1900. The relatively small German ancestry fraction may be due to the restrictions that the Moravian brethren put on intermarriage with the Greenlandic population were efficient.

The inflow of Danes to Greenland particularly from the 1940s marked a substantial increase in immigration rate of Europeans to Greenland, very likely contributing to the predominantly Danish source of European ancestry. This is also well in line with the many individuals in our study that we infer to have at least one entirely European parent. Furthermore, among the Greenlandic individuals that do not have traces of European admixture, most live in the very North as well as the East coast of Greenland as shown in previous research (Moltke *et al.*, 2015). This aligns well with Danish colonial activities that were initiated later in the North (1909) and East (1894) than in the Southwest (1721). Concurrently, it seems reasonable to argue that Danish colonial trade and administrative policies have been an important factor in shaping the genetic composition of the current day Greenlandic population. Taken together, these results seem consistent with recent demographic trends in Greenland and with historical records of Europe contact.

References

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Bjerregaard P (2011). *Inuit Health in Transition: Greenland Survey 2005-2010: Population Sample and Survey Methods*. Statens Institut for Folkesundhed.

Bjerregaard P, Curtis T, Borch-Johnsen K, Mulvad G, Becker U, Andersen S, *et al.* (2003). Inuit health in greenland: a population survey of life style and disease in greenland and among inuit living in denmark. *Int J Circumpolar Health* **62 Suppl 1**: 3–79.

Bosch E, Calafell F, Rosser ZH, Nørby S, Lynnerup N, Hurles ME, *et al.* (2003). High level of male-biased scandinavian admixture in greenlandic inuit shown by y-chromosomal analysis. *Hum Genet* **112**: 353–363.

Brown RNR (1951). Whaling: Greenland and davis strait fishery. Unpublished.

Chacón-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, *et al.* (2018). Latin americans show wide-spread converso ancestry and imprint of local native ancestry on physical appearance. *Nat Commun* **9**: 5388.

Daveluy M, Lévesque F, Ferguson J (2011). Humanizing security in the arctic. `https://www.uap.ualberta.ca/titles/119-9781896445540-humanizing-security-in-the-arctic`. Accessed: 2019-3-5.

Delaneau O, Zagury JF, Marchini J (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**: 5–6.

Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, *et al.* (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**: 295–302.

Ebenesersdóttir SS, Sandoval-Velasco M, Gunnarsdóttir ED, Jagadeesan A, Guðmundsdóttir VB, Thordardóttir EL, *et al.* (2018). Ancient genomes from iceland reveal the making of a human population. *Science* **360**: 1028–1032.

Egedes H (2017). *Henriette Egedes Dagbog 1832-1833*. Edited by Inge Høst Seiding, Det Grønlandske Selskab, Denmark.

Frandsen N (2010). *Nordgrønland 1790-96 - inspektør B.J. Schultz' indberetninger til direktionen for den Kongelige grønlandske Handel*. Selskabet for Udgivelse af Kilder til Dansk Historie.

Frandsen N, H GH, C HJ, Jensen EL, Marquardt O, Rud S, *et al.* (2017). Grønland – den arktiske koloni. In: Gulløv HC (ed.), *Danmark og kolonierne*, Gads Forlag, København, pp. 46–107.

Friesen TM, Arnold CD (2008). The timing of the thule migration: New dates from the western canadian arctic. *Am Antiq* **73**: 527–538.

Gad F (1969). *Grønlands historie, II, 1700-1782*. Nyt Nordisk Forlag Arnold Busck, Denmark.

Gelman A, Rubin DB (1992). Inference from iterative simulation using multiple sequences. *Stat Sci* **7**: 457–472.

Gulløv HC (2008). The nature of contact between native greenlanders and norse. *Journal of the North Atlantic* pp. 16–24.

Hamilton LC, Rasmussen RO (2010). Population, sex ratios and development in greenland. *Arctic* **63**: 43–52.

Hansen P (1955). De fremmede fiskere i vestgrønlands farvande. *Tidsskriftet Grønland* **5**: 190–200.

Hansen P (1961). De senere års fiskeri ved vestgrønland. *Tidsskriftet Grønland* **10**: 361–370.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, *et al.* (2014). A genetic atlas of human admixture history. *Science* **343**: 747–751.

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

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, *et al.* (2015). The fine-scale genetic structure of the british population. *Nature* **519**: 309.

Maples BK, Gravel S, Kenny EE, Bustamante CD (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**: 278–288.

Meisner J, Albrechtsen A (2019). Testing for hardy-weinberg equilibrium in structured populations using genotype or low-depth ngs data. *Molecular Ecology Resources* **0**. **URL:** *https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13019*

Moltke I, Albrechtsen A (2014). RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics* **30**: 1027–1028.

Moltke I, Fumagalli M, Korneliussen TS, Crawford JE, Bjerregaard P, Jørgensen ME, *et al.* (2015). Uncovering the genetic history of the present-day greenlandic population. *Am J Hum Genet* **96**: 54–69.

Ostermann H (1940). *Nordmænd på Grønland 1721–1814*, vol. 1. Gyldendal Norsk Forlag.

Pereira V, Tomas C, Sanchez JJ, Syndercombe-Court D, Amorim A, Gusmão L, *et al.* (2015). The peopling of greenland: further insights from the analysis of genetic diversity using autosomal and x-chromosomal markers. *Eur J Hum Genet* **23**: 245–251.

Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, *et al.* (2014). The genetic prehistory of the new world arctic. *Science* **345**: 1255832.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, *et al.* (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**: 757–762.

Seiding I (2013). "married to the daughters of the country': Intermarriage and intimacy in northwest greenland ca. 1750 to 1850". Ph.D. thesis, University of Greenland.

Seiding I (2016). *Listevis af liv. Grønland i Tal - Kilder og cases gennem 300 år*. Aarhus University Press.

Seiding IH (2011). Intermarriage in colonial greenland 1750–1850: Governing across the colonial divide. In: Daveluy M, Lévesque F, Ferguson J (eds.), *Humanizing Security in the Arctic*, CCI Press.

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, *et al.* (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* **8**: e1002793.

Wilhjelm H (2001). *Af tilbøielighed er jeg grønlandsk. Om Samuel Kleinschmidts liv og værk*, *Det Grønlandske Selskab skrifter*, vol. 34. Det Grønlandske Selskab.

Supplemental Tables

CHR	cM	nSNPs	mean ancestry chunks	means SNPs per chunk
1	292.1	10398	524.4	19.8
2	274.3	11109	502.4	22.1
3	227.1	9419	428	22
4	219.4	8233	402.4	20.5
5	208.6	7976	386.8	20.6
6	197.8	9336	367.4	25.4
7	189.5	7774	354.2	21.9
8	177.4	7189	312.7	23
9	179.4	6599	311.4	21.2
10	182.2	7070	331.7	21.3
11	161.5	6685	306.1	21.8
12	173.9	6697	320.3	20.9
13	128.3	4915	241.9	20.3
14	115.5	4430	224.2	19.8
15	150.8	4346	233.8	18.6
16	130.7	4496	246.4	18.2
17	127.9	4039	243.5	16.6
18	119.7	3907	224.5	17.4
19	106.6	2914	201.7	14.4
20	109.8	3698	207.8	17.8
21	63.5	2251	120.8	18.6
22	72.4	2221	133.7	16.6

Supplemental Table 1. Summary of CHROMOPAINTER results from each chromosome showing the number of SNPs, and the number of distinct ancestry chunks on each chromosome. Chunk values are averages across all analyzed individuals, including both Europeans and Greenlanders.

	min1%			min5%			min20%		
Threshold	0.95	0.99	0.999	0.95	0.99	0.999	0.95	0.99	0.999
Belgium	1	0	0	0	0	0	0	0	0
Denmark	1287	1208	1109	1190	1100	1002	639	567	483
Dutch	3	1	1	1	1	1	1	1	1
Finland	11	9	7	5	5	5	1	1	1
France	0	0	0	0	0	0	0	0	0
Greenlandic Inuit	1555	1555	1555	1555	1555	1555	1542	1541	1540
Germany	10	1	0	4	1	0	0	0	0
Ireland	6	3	3	4	3	0	0	0	0
Italy	0	0	0	0	0	0	0	0	0
Northern Ireland	1	0	0	0	0	0	0	0	0
Norway	364	281	210	133	98	75	20	18	17
Poland	8	2	0	3	1	0	0	0	0
Spain	0	0	0	0	0	0	0	0	0
Sweden	90	56	34	31	20	13	3	1	1
UK	8	5	3	5	3	3	1	1	1

Supplemental Table 2 Assignment to country at 1%, 5% and 20% ancestry thresholds across 1582 admixed Greenlanders. The values shown are counts of the number of individuals inferred to have at least [1%, 5%, 20%] ancestry, arranged by column from each source country, as in Table 1 in the main text. Sub-columns give counts at different support thresholds.

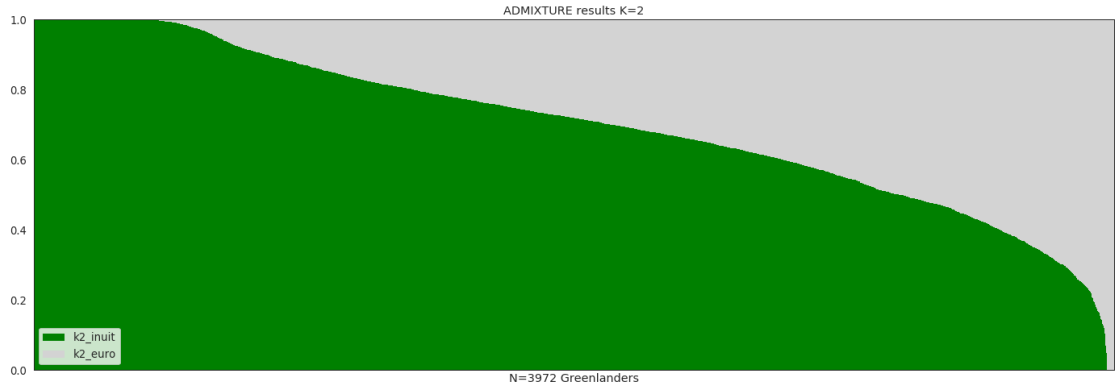
	min1%			min5%			min20%		
Theshold	**0.95**	**0.99**	**0.999**	**0.95**	**0.99**	**0.999**	**0.95**	**0.99**	**0.999**
Belgium	1	0	0	0	0	0	0	0	0
Denmark	1293	1218	1122	1204	1122	1024	664	583	490
Dutch	3	1	1	1	1	1	1	1	1
Finland	11	9	7	5	5	5	1	1	1
France	0	0	0	0	0	0	0	0	0
Greenlandic Inuit	1555	1555	1555	1555	1555	1555	1542	1541	1540
Germany	19	1	1	10	1	0	1	0	0
Ireland	6	3	3	4	3	0	0	0	0
Italy	0	0	0	0	0	0	0	0	0
Northern Ireland	1	0	0	0	0	0	0	0	0
Norway	357	277	205	132	96	77	20	18	17
Poland	6	2	2	3	2	0	0	0	0
Spain	0	0	0	0	0	0	0	0	0
Sweden	93	58	36	38	20	13	3	1	1
UK	8	5	3	6	4	3	1	1	1

Supplemental Table 3. Assignment to country at 1%, 5% and 20% ancestry thresholds across 1582 admixed Greenlanders, with sparse prior. The values shown are counts of the number of individuals inferred to have at least [1%, 5%, 20%] ancestry, arranged by column from each source country, as in Table 1 in the main text. Sub-columns give counts at different support thresholds.

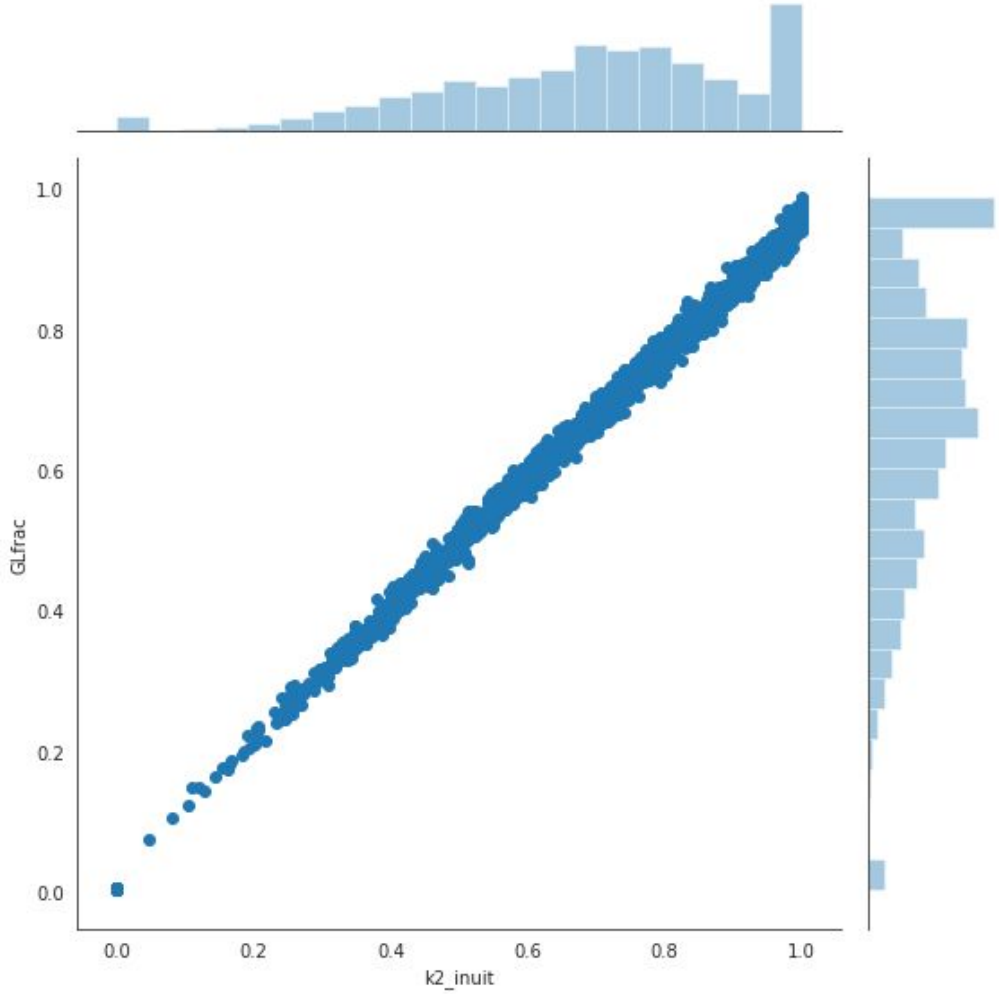
	Group-based
	SOURCEFIND
Belgium	0.1%
Denmark	32.3%
Dutch	0.0%
Finland	0.0%
France	0.2%
Greenlandic Inuit	65.6%
Germany	0.3%
Ireland	0.1%
Italy	0.1%
Northern Ireland	0.1%
Norway	0.6%
Poland	0.1%
Spain	0.1%
Sweden	0.2%
UK	0.1%

Supplemental Table 4. Effect of a sparse prior on the grouped SOURCEFIND analysis. In the group-based analysis (right) the values shown are the percentage ancestry inferred to to come from each country, by SOURCEFIND and NNLS. Compare to Table 1 in the main text.

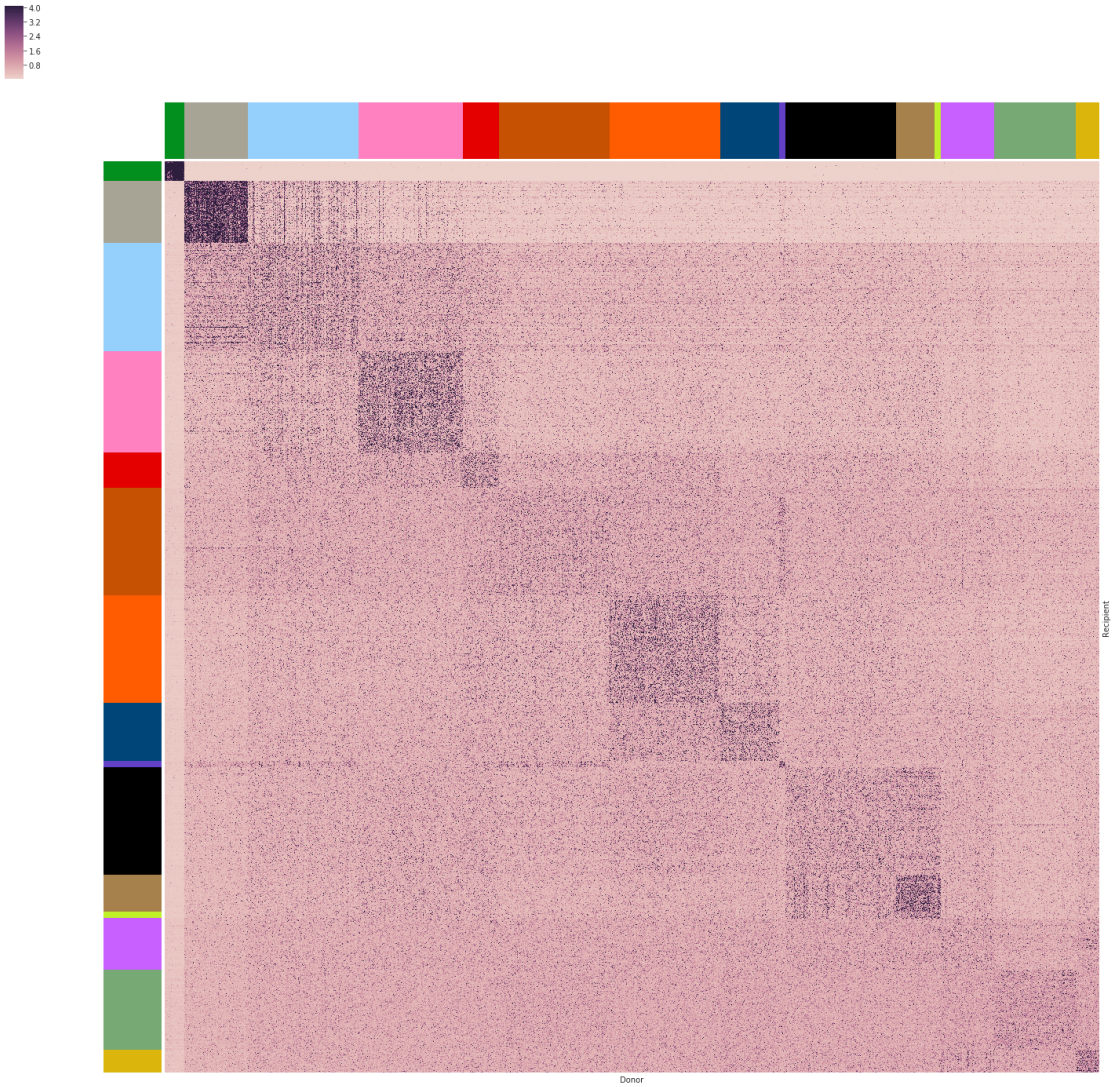
Supplemental Figures



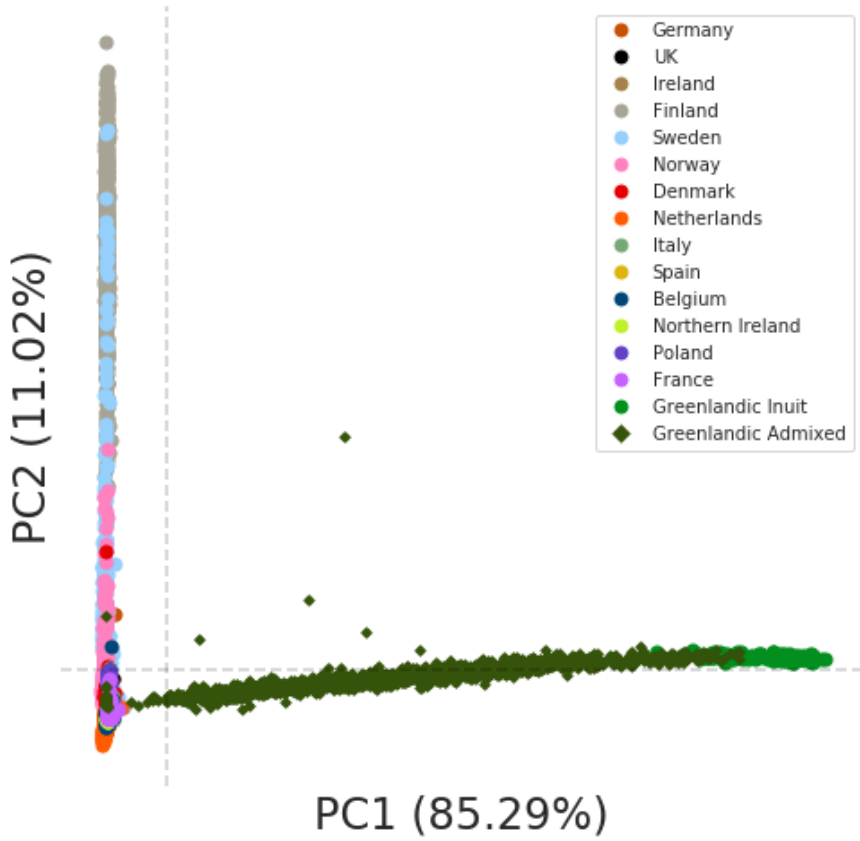
Supplemental Figure 1. Barplot of the unsupervised ADMIXTURE analysis (K=2) for 3972 Greenlanders. The Greenlandic Inuit ancestry is shaded green. This analysis was used to identify admixed and non-admixed Greenlander for downstream analyses.



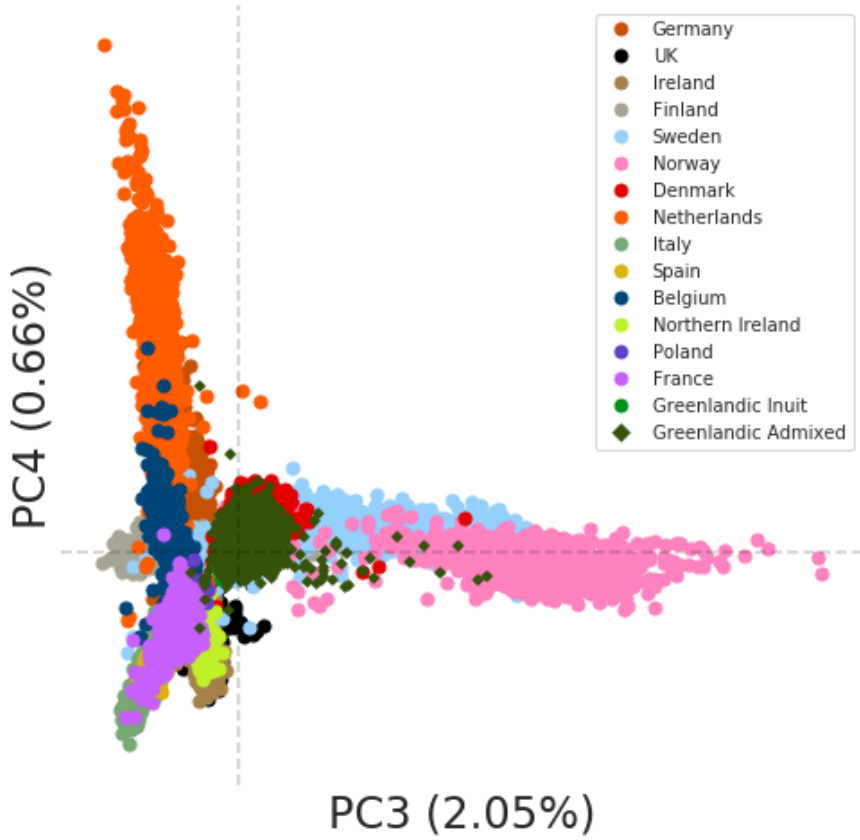
Supplemental Figure 2. Agreement between the ADMIXTURE K=2 Greenlandic Inuit ancestry proportion in each admixed Greenlandic individual (see Supplemental Figure 1) and the proportion of the genome copied from reference non-admixed Greenlanders in CHROMOPAINTER. The correlation coefficient for this relationship is 0.9976.



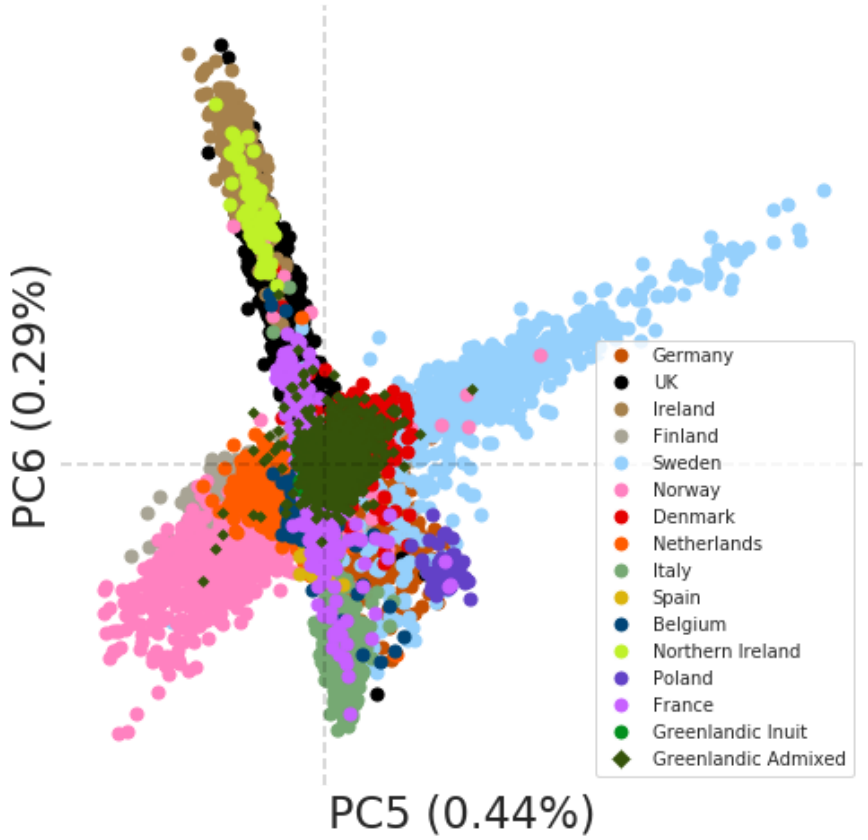
Supplemental Figure 3. Coancestry matrix, between 8456 reference individuals, as estimated by CHROMOPAINTER, based on “chunk lengths”. Colors along the axes show the country of of origin for each individual. Individuals are not ordered within each country.



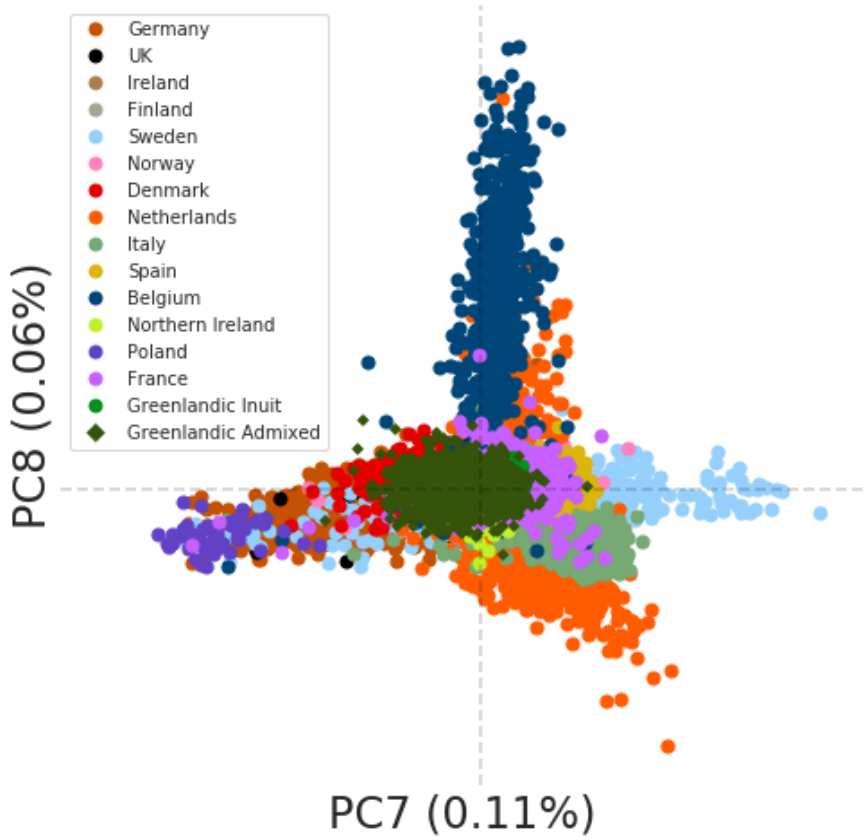
Supplemental Figure 4. Scatter plots of principal component (PC) axes. Colored shapes represent reference (circle) and admixed (diamond) individuals, and are shaded by country. The percentages given on each axis show the percentages of variance explained. PC analysis conducted on the CHROMOPAINTER coancestry matrix, summed over the reference individuals from each country. Plotting z-order is the same as the legend.
(Continued below)



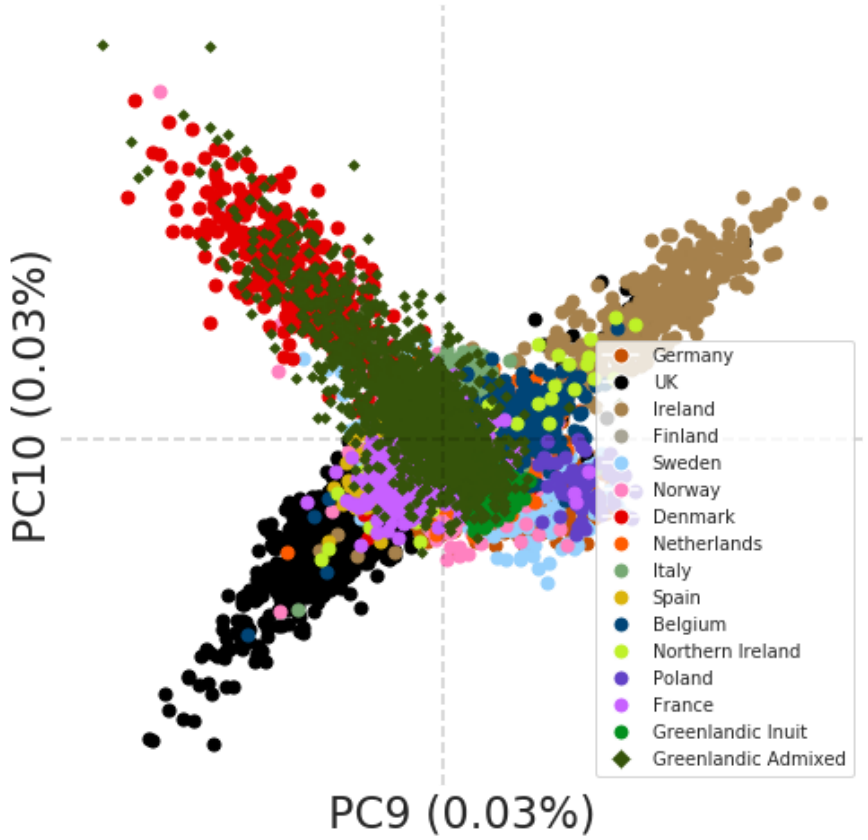
Supplemental Figure 4 (continued). Scatter plots of principal component (PC) axes.



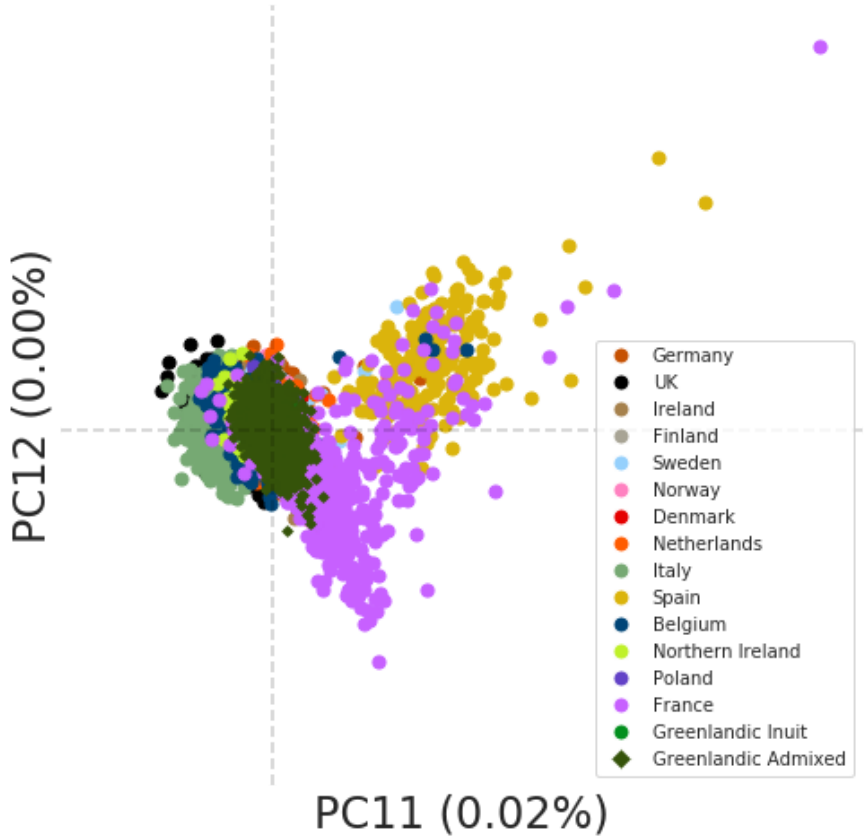
Supplemental Figure 4 (continued). Scatter plots of principal component (PC) axes.



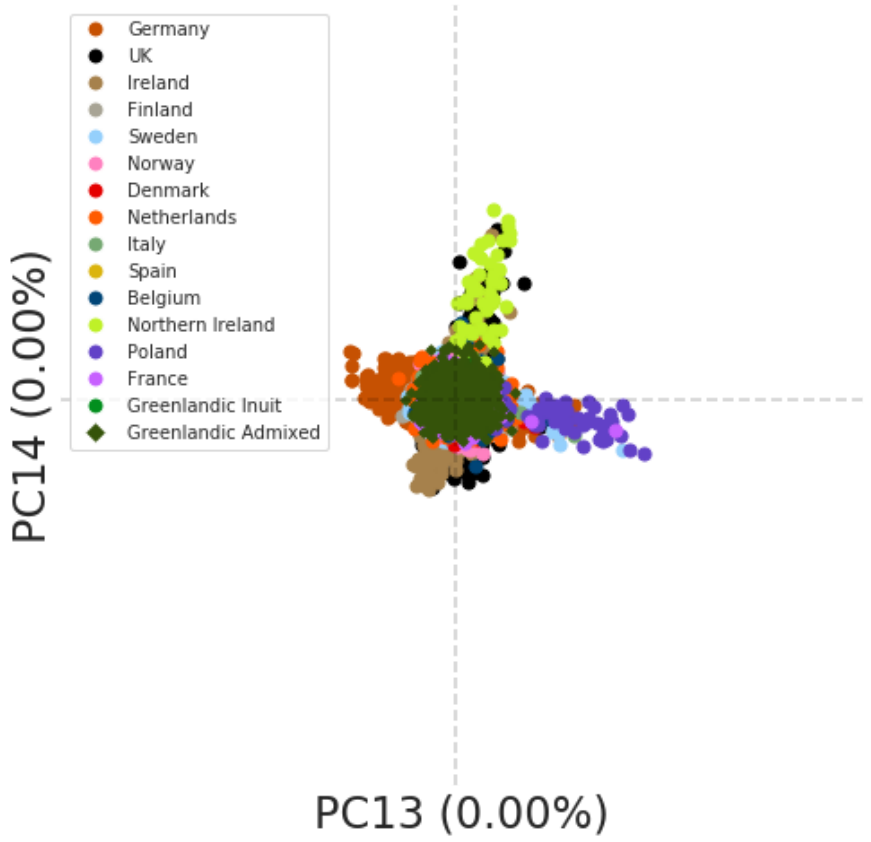
Supplemental Figure 4 (continued). Scatter plots of principal component (PC) axes.



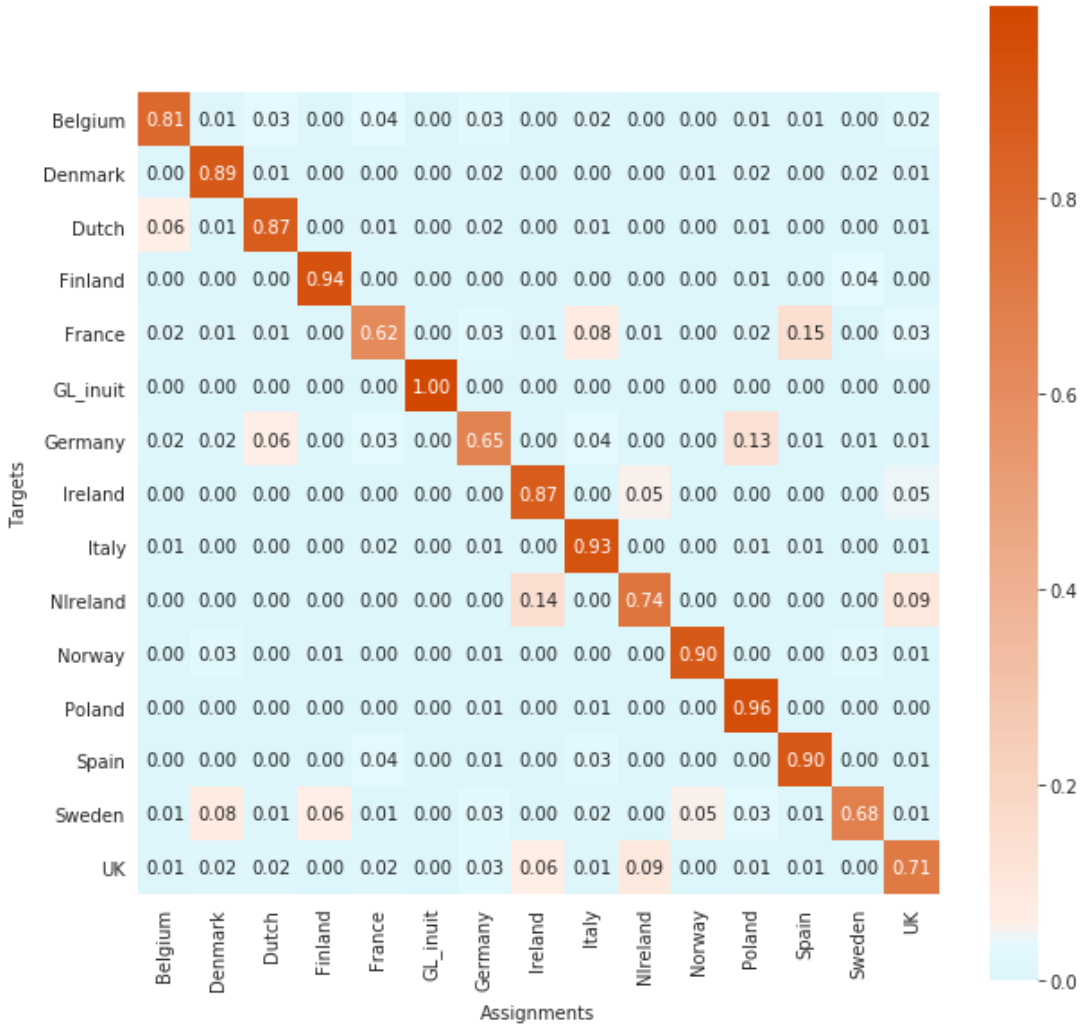
Supplemental Figure 4 (continued). Scatter plots of principal component (PC) axes.



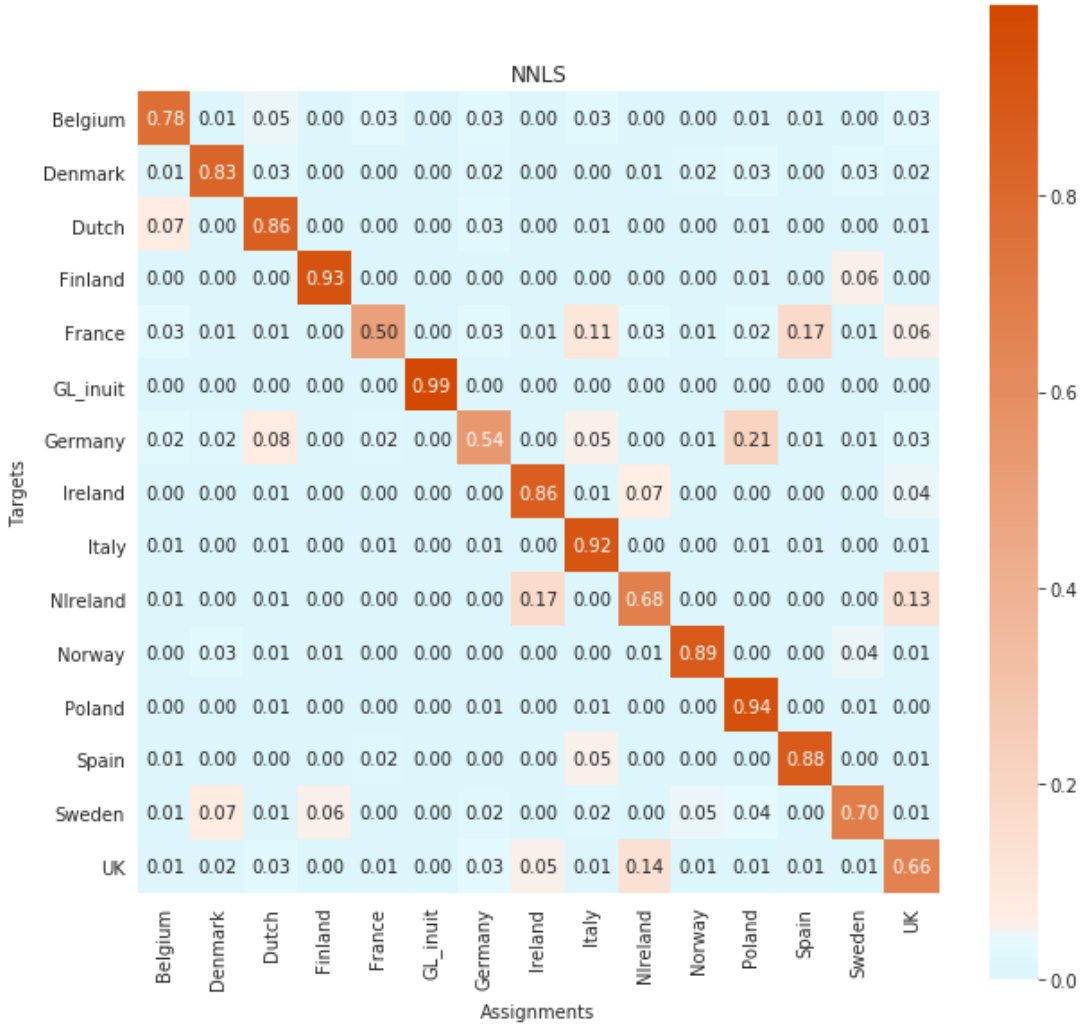
Supplemental Figure 4 (continued). Scatter plots of principal component (PC) axes.



Supplemental Figure 4 (continued). Scatter plots of principal component (PC) axes. Colored shapes represent reference (circle) and admixed (diamond) individuals, and are shaded by country. The percentages given on each axis show the percentages of variance explained. PC analysis conducted on the CHROMOPAINTER coancestry matrix, summed over the reference individuals from each country. Plotting z-order is the same as the legend.



Supplemental Figure 5. The results of a leave-one-out SOURCEFIND analysis. The numbers indicates the mean ancestry proportion estimate for each reference country obtained by analysing one reference sample at a time while leaving it out and using the remaining samples as references.



Supplemental Figure 6. Assignment proportions from the leave-one-out analysis with NNLS. The number in each cell indicates the mean ancestry proportion estimate for each reference country obtained by analysing one reference sample at a time while leaving it out and using the remaining samples as references. Compared to the SOURCEFIND analysis (above), the self-assignment rates are slightly lower.

Paper III:
Estimating linkage disequilibrium in admixed populations

By

Ryan K. Waples[1], Anders Albrechtsen[1], Ida Moltke[1]

1 The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Publication details

In preparation

ESTIMATING LINKAGE DISEQUILIBRIUM IN ADMIXED POPULATIONS

(IN PREPARATION)

Ryan K. Waples, Anders Albrechtsen, Ida Moltke

Section for Computational and RNA Biology,
Department of Biology, University of
Copenhagen, Copenhagen N, Denmark

ABSTRACT

Analysis of linkage disequilibrium (LD) is key to understanding many aspects of population genetics. The LD in a population is affected by many aspects of the populations history including past population size and admixture. Due to these effects, current LD patterns are informative about the past and present of populations. However, in many populations, recent admixture has a large impact on LD patterns, potentially masking the ancestry-specific LD that existed prior to admixture.

We present LDadmix, a tool to estimate two-locus haplotype frequencies within the source ancestries of a recently admixed population, with the goal of recovering the LD patterns within each ancestry source prior to the admixture event. It requires only unphased genotype data and estimates of admixture proportions for a set of samples from an admixed population. We apply LDamix to simulated data, as well as data from admixed human populations in the Americas, to assess its accuracy and demonstrate its utility by estimating r^2 decay curves for each ancestry of a recently admixed population. In the Americas, we recover an elevated LD decay curve for the ancestral American ancestry, after accounting for recent African and European admixture.

LDadmix is available as an open source Python program hosted at `https://github.com/rwaples/LDadmix`. The software uses common data file formats (bed/bim/fam) and is suitable for analysis on data set with thousands of individuals and millions of pairs of loci.

Introduction

Linkage disequilibrium (LD) is the nonrandom association between alleles at different genomic loci. Across the genome, LD is driven by a combination of local genomic factors e.g., sequence motifs guiding recombination (Grey, Baudat, and de Massy 2018), and selection (Kawakami et al. 2014) as well as demographic factors such as genetic drift and admixture that have a genome-wide impact (Slatkin 2008). LD plays a vital role in the study design of many different types of investigations, as it provides information about the statistical (Pritchard and Przeworski 2001) and genealogical (McVean 2002) independence of genetic variation across the genome.

LD patterns can also help us understand the past and present of a population, since the LD patterns of a population have been shaped by its demographic history (Pritchard and Przeworski 2001). For example, LD across different genomic distances can be used to make inferences about past population size changes (Tenesa et al. 2007; Myers, Fefferman, and

Patterson 2008; Ragsdale and Gutenkunst 2017), and recent effective population size (N_e) (Hill 1981; Waples and Do 2010). Broadly, the methods for performing such inference relate observed LD patterns to a time series of N_e values by balancing the effects of genetic drift, which can create LD, and recombination, which partially breaks down exisiting LD each generation. Even simple summaries of LD, like mean r^2 for pairs of loci at different distances, here called a LD decay curve (Sved 1971; W. G. Hill and Weir 1988), contain a lot of information about a population's history, with higher r^2 signalling lower N_e. Comparing the LD decay curves of two or more populations can be a way to see if they share similar demographic histories. However, caution is required with small sample sizes, as they induce a bias in r^2 of approx. $1/n$, with sample size n (Weir and Hill 1980).

In a similar manner, LD, and LD decay curves, can help us learn about admixture in both archaic (Plagnol and Wall 2006; Ragsdale and Gravel 2018) and more recent timescales (Moorjaniet al. 2013; Loh et al. 2013; Hellenthal et al. 2014). This is very important because, beyond the recent past, admixture seems nearly universal among human populations (Hellenthal et al. 2014) and is common and or underestimated in many other species (Supple and Shapiro 2018). When two or more populations mix, the process of admixture affects LD, with the LD post-admixture a function of the LD within each admixing population, the allele frequency differences between populations, the admixture proportions, and time since admixture (Chakraborty and Weiss 1988).

The effect of demography and admixture on LD is illustrated in Figure 1. We see that a population that has gone through a bottleneck (simulated East Asian) has an LD decay curve that is higher and flatter at short genetic distances, compared to a population that has not (simulated West African), with lower LD and a steeper decay at short distances. Furthermore, in the LD decay curve for the population that is an admixture between the two other populations, we see a transition between two LD regimes. At shorter distances, where there is considerable background LD within each source population, the admixed population has intermediate r^2 values. At longer distances, there is little background LD present in the source populations, and the admixed population has more LD than either source population.

Unfortunately, due to its effect on LD, admixture can obscure the ancestry-specific LD that existed prior to admixture (e.g. Moltke et al. 2015), and with only access to samples from an admixed population, the pre-admixture LD within each source population is difficult to recover. In turn, this means that it is difficult to recover ancestry-specific demographic factors, and complicates the interpretation of LD decay curves in admixed populations.

The confounding of admixture-induced LD with LD that existed prior to admixture is evident in several real populations. The LD decay of the Greenlandic Inuit, a population isolate with extensive LD and significant European admixture is very different when measured across a random sample of the population versus a set of unadmixed individuals of Inuit descent (see Figure 6 of Moltke et al (2015)). There are also admixed population samples in the 1000 Genomes (1000G) (1000 Genomes Project Consortium et al. 2015). In fact, the Native American populations included in the 1000G are all admixed, making it difficult to interpret their LD decay curves (e.g. extended data figure 10 of 1000 Genomes Project Consortium et al. 2015). As a group, their LD decay curves are both above and below those of European and East Asian populations, complicating a demographic interpretation. Based on such data is it unclear to what extent the pattern of LD decay reflect the demographic history of the Native American ancestry, and to what extent do they reflect the recent history of admixture with African and European populations.

Here we present software to estimate two-locus haplotype frequencies within the source populations of recently admixed populations, based on a model first presented in Moltke et al (2015), and developed further here. This method takes as input genotypes and admixture proportions for a set of individuals and produces estimated frequencies for the four possible two-locus haplotypes within each admixture component. The process to estimate these haplotype frequencies from a set of samples is 1) genotype the samples at a set of loci, 2) run ADMIXTURE (Alexander, Novembre, and Lange 2009) or a similar program to fit an admixture model with appropriate K, and 3) run LDadmix to estimate the haplotype frequencies within each of the K ancestries. Established two-locus LD measures such as r^2 and D are also calculated for each ancestry by LDadmix.

(WAPLES ET AL., IN PREPARATION)

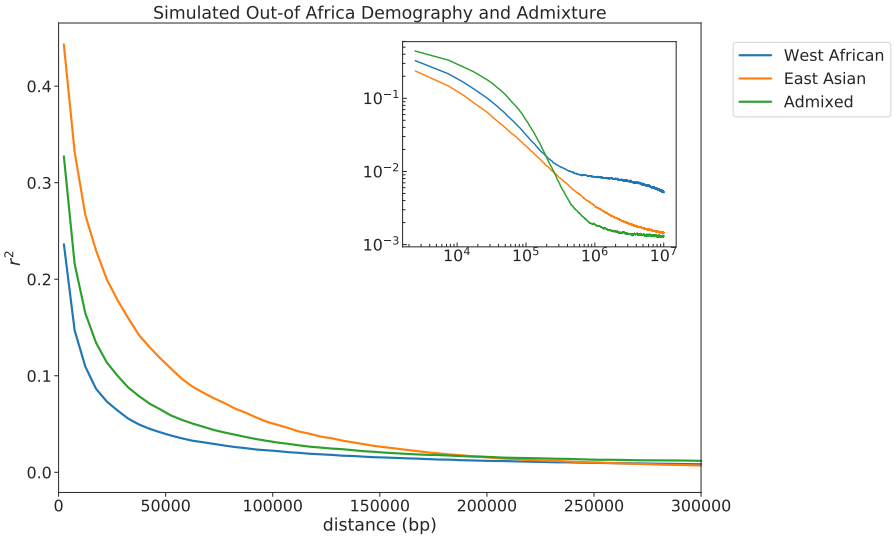


Figure 1: Illustration of some effects of demography and admixture on LD decay curves. LD decay curves for a simulation of an admixed population and its two source populations. One of these source populations (East Asian) has been simulated to have gone through an out-of-Africa like bottleneck and one that has not (West African). Each line shows r^2 decay in a single population with increasing distance between loci. The green and orange lines show LD in the source populations, while the blue line shows LD in a population that is an mixture of the two source populations. Both the large plot and the inset present the same data, shown on different scales. The large plot shows distance up to 300Kb, beyond which the background LD in each source population is minimal; the inset plot is log scaled on both axes and extends the distance to 10M bp. The haplotypes and admixture were simulated using the coalescent and an out-of-Africa demographic model; 800 haplotypes were sampled from each population. See methods for a full description of the simulation procedure.

The data requirements for LDadmix are much lower than alternative methods to estimate ancestry specific-LD from admixed samples. We are not aware of any other method to accomplish this directly, but one possible method is to mask ancestry tracts from non-target ancestries via local ancestry inference. Unfortunately, the best-performing methods of local ancestry inference (e.g. Maples et al. 2013) require phasing as well as reference haplotypes from each possible ancestry. LDadmix requires neither phased data or ancestry references, and is likely most useful when there is no large set of reference haplotypes or unadmixed individuals available.

Below we apply LDadmix to estimate LD decay curves in both simulated and real data sets. First, we examine its accuracy on simulated data, and then we shift to an analysis of admixed human populations from the Americas. In the Americas, we attempt to recover the ancestral pattern of LD in the Native American ancestry component while accounting for recent admixture with African and European ancestry components.

Materials & Methods

Model

Following Moltke et al. (2015) we assume we have genotype data G for n individuals from a population that is an admixture of K source populations and take the the likelihood of the haplotype frequencies in the source populations p as:

$$P(G|p,\alpha) = \prod_{i=1}^{n} \sum_{h \in h(G_i)} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p_{h_1}^{k_1} p_{h_2}^{k_2} \alpha_i^{k_1} \alpha_i^{k_2}. \quad (1)$$

Here p_j^k denotes the frequency of haplotype j for the kth population, α_i^k denotes the admixture proportions of individual i in the kth population, and $h = (h_1, h_2)$ is the unobserved pair of haplotypes for an individual, with the two haplotypes originating from the unobserved ancestral populations k_1 and k_2, respectively. Finally, the term $h(G_i)$ denotes the set of pairs of haplotypes that are consistent with the observed genotype.

We use an expectation-maximization (EM) algorithm to find the haplotype frequencies that maximize this likelihood, by alternately updating the estimates of the haplotype frequencies (M-step) and the contributions of each individual given their admixture proportions and genotypes (E-step). We start the EM with a random initialization of haplotype frequencies and continue until the change in likelihood falls below a threshold value (default = 1e-6). This formulation provides a natural extension to any number of admixing populations and with only a single population, reduces to the method of Excoffier and Slatkin (1995).

Note that this model is designed to be applied to pairs of loci that are relatively close in the genome; we assume that each two-locus haplotype originates from a single ancestral population, with the same ancestry at both sites. This does not account for any recombination that has occurred since admixture that could generate two-locus haplotypes with a distinct ancestry at each site. The assumption of no recombination is unlikely to be strictly true, but is reasonable as long as the admixture is recent and the loci are unlikely to recombine in each generation. Low levels of recombination are unlikely to drastically alter haplotype frequencies, and so have a limited effect on most LD measures at small genomic distances. However, the effect of limited recombination will depends on the choice of LD measure, and will also have the largest effect on LD with rare alleles.

Simulated Data

Simulation of out-of-Africa scenario with admixture. To produce the example admixture scenario illustrated in Figure 1, we generated samples from an out of Africa demography (Gutenkunst et al. 2009) using msprime (v7.0) (Kelleher, Etheridge, and McVean 2016), setting background migration rates to zero for the most recent 100 generations. Specifically we sampled from a West African-like population and an East Asian-like population, as well as a novel admixed population, created in a single pulse admixture event three generations prior to sampling, with 30% ancestry from West Africa and 70% from East Asia. We simulated a 100M bp chromosome, with recombination and mutation rates both equal to 1e-8 per bp. From each source population and the admixed population ("West African", "East Asian", "Admixed"), 800 haplotypes were sampled and for each set of haplotypes we applied a 5% minor allele frequency (MAF) filter, and selected 50K variable sites at random. The haplotypes from each population were used to calculate mean r^2 in 5kb distance bins. For better illustration of the extent of admixture LD, we extended analysis to locus pairs separated by up to 10M bp.

Simulation to assess LDadmix. Simulated data to assess LDadmix were generated by sampling two-locus haplotypes from 1000 Genomes phased data. To construct data for an admixed population consisting of 200 individuals we sampled from haplotypes present in 100 individuals from each of the West African Yoruban Nigerian population (YRI) and the Chinese Han from Beijing (CHB) population samples. We treat each pair of loci separately, using the 200 two-locus haplotypes from each population to construct an admixed population of 200 diploid individuals, utilizing all haplotypes. The simulated admixed individuals were assigned haplotypes from each population based on their admixture proportions.

We assessed the performance of LDadmix on two different distributions of admixture proportions, representing distinct admixture scenarios. In each case the mean expected ancestry proportion from each population was 0.5, but the cases differed in how the ancestry was distributed across individuals. In the first case, admixture proportions for the first

population were sampled from a high variance beta distribution (α = 0.1, β =0.1) (Figure 2A). In the the second case, admixture proportions for the first population were sampled from a uniform (0,1) distribution (Figure 2B). Simulated individuals were assigned an admixture proportion for each scenario that was held constant across all pairs of loci. The two-locus haplotypes from YRI and CHB were assigned to the simulated individuals using a rejection sampling procedure to ensure that all 200 haplotypes from each source population were utilized. For each individual, the number of haplotypes from the first population was selected based on a two draw binomial distribution parameterized with the admixture proportions described above.

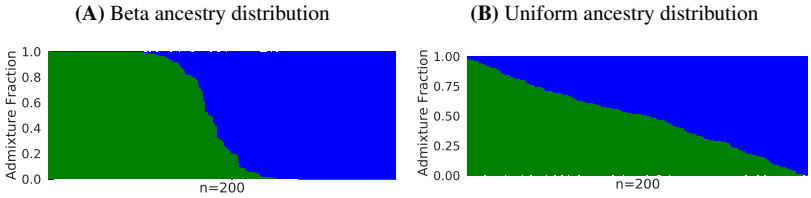


Figure 2: Bar plots of the admixture proportions in the two simulation scenarios. Each plot shows the admixture proportions for 200 individuals, drawn from a distribution with expected proportion of the first population of 0.5: **A)** ancestry proportions drawn from a beta(0.1, 0.1), **B)** ancestry proportions drawn from a uniform (0,1).

We used phased data from 100K SNP sites randomly selected from chromosome 18. Starting from the 1000 Genomes phased vcf files, we selected di-allelic sites with a minor allele count of at least two across all populations. From this set of SNPs, we selected 100K sites at random. By using real data as the basis for simulated admixture, we capture some of the complexities present in the LD of real populations while also retaining the ability to have a true sample LD to compare against.

In addition, we applied LDadmix to test a few additional cases where it should be more difficult or easier to resolve ancestry-specific LD patterns. Specifically, we evaluated performance under a more even ancestry proportion distribution, given by a triangle distribution with mode=0.5, min=0.25 and max=0.75 (Supplementary Figure S2). This distribution provides much less information about the ancestry source of each haplotype than the alternatives tested above, and so should be a difficult test for LDadmix. We also evaluated the effect of supplementing with additional unadmixed samples from one of the source populations: after the rejection sampling procedure for each pair of loci, we supplemented the data set supplied to LDadmix with genetic data from eight unadmixed YRI individuals. This provides additional information about the haplotype frequencies for the YRI ancestry component and should therefore aid in LD estimation.

1000 Genomes data

We selected four admixed American population samples with significant Native American ancestry: PEL - Peruvians from Lima, Peru, CLM - Colombians from Medellin, Colombia, PUR - Puerto Ricans from Puerto Rico, and MXL - Mexican Ancestry from Los Angeles USA. We combined these population samples with 11 other population samples selected to include African, European, East Asian, and Native American ancestries: ACB, ASW, LWK, YRI, CEU, IBS, GBR, FIN, CHB, CDX, CDX, CHS. We fit a K=4 admixture model on this set of individuals with ADMIXTURE (Alexander, Novembre, and Lange 2009). Using the results of this admixture model, we selected a reduced set of individuals where a K=3 admixture model seemed reasonable by excluding the CDX, CHS, and CHB individuals, the FIN individuals, the ACB and AWK population samples, as well as seven American individuals estimated to have more than 5% Asian ancestry. On this set of 844 individuals from 9 population samples, we fit a K=3 admixture model with ADMIXTURE, and used the admixture proportions estimated here and genotype data as input to LDadmix.

We also constructed six further datasets by excluding groups of individuals from the above K=3 dataset to examine to what extent they were driving the results. To make these data sets, we excluded individuals in two ways: 1) dropping all American individuals except those from each of the four America populations in turn: PEL, MXL, PUR, CLM, (4 sets)

and 2) dropping individuals from the population with the most Native American ancestry, PEL, either all of them, or just the 18 samples with more 99% Native American ancestry (2 sets). We applied LDadmix to each of these data sets using the admixture proportions estimated from the full K=3 dataset.

To construct genotype data sets for the above analyses, we first extracted eligible di-allelic sites from the 1000G phase 3 vcf files aligned to GRCh37 (available at http://www.internationalgenome.org/). Then for each LDadmix analysis, we selected the appropriate individuals, applied a minor allele count filter of 2, and took 500K sites at random. To construct a dataset for each ADMIXTURE analysis, we selected the appropriate individuals, applied a minor allele frequency filter of 0.05 to the eligible sites, and took 500K sites at random. We ran ten replicates of each ADMIXTURE analysis, selecting the run with best likelihood, and checking for convergence by checking for multiple other runs within 2 log likelihood units.

LD decay estimation

For both the simulated and real data sets we applied LDadmix to all pairs of loci separated by up to 300kb, stopping the EM when the log likelihood changed by less than 1e-6. From the estimated haplotype frequencies, we calculated three LD measures (r^2, D, D') within each ancestry, but focused on r^2.

For each data set we also calculated LD with existing methods for comparison. For the simulated data, we calculated 'true' LD values from the phased haplotypes from each population, representing the LD present in this sample. For the real data, we estimated LD within each population sample separately using the method of Gaunt, Rodríguez, and Day (2007) as implemented in PLINK (v1.9) (Chang et al. 2015), which produces very similar estimates to the EM model above to a single-population (data not shown).

When summarizing LD, we restrict the analyses to pairs of loci where both sites are estimated to have MAF >0.05. For ancestral populations, we estimated the MAF from the haplotype frequencies estimated by LDadmix. The goal of the MAF filter filter was to ensure that only variable sites are considered within each ancestry, and also to reduce the effect of rare alleles.

To generate r^2 decay curves, we calculated a 5kb unweighted moving average over distances, assigning the mean r^2 value for each window to the midpoint, so the decay curve covers the distance from 2.5kb to 297.5kb.

For the simulated data, we calculated the deviation between the true and estimated r^2 values with two different measures: root-mean-square deviation (RMSD) and mean bias. We calculated these deviations across the entire data set and in non-overlapping 5kb distance windows.

Results

Application to simulated data

First, we examined the accuracy of LDadmix by applying it to simulated data. To generate the simulated data sets, we sampled existing two-locus haplotypes from two population samples from the 1000 Genomes; YRI and CHB, and combined them into an admixed diploid population. Haplotypes were assigned to individuals based on the admixture fraction (Q) assigned to each individual, using a rejection sampling procedure (see methods). To explore the effect of the Qs, we used two different sets of Qs, both with a mean ancestry of 0.5: a set of Qs sampled from a higher variance beta distribution (α = 0.1, β=0.1) and a set of Qs sampled from a lower variance uniform distribution, both shown in Figure 2. The LD decay curves for each of the YRI and CHB source population samples, as well as the simulated admixed data are shown in Supplemental Figure S1.

We applied LDadmix to the two simulated admixed datasets using the known sets of Qs and obtained LD estimates for 34M pairs of loci separated by less than 300kb for each of the two ancestry source populations. We here report r^2 for approx. 3M pairs of loci within each ancestry with both loci estimated to have minor allele frequency (MAF)

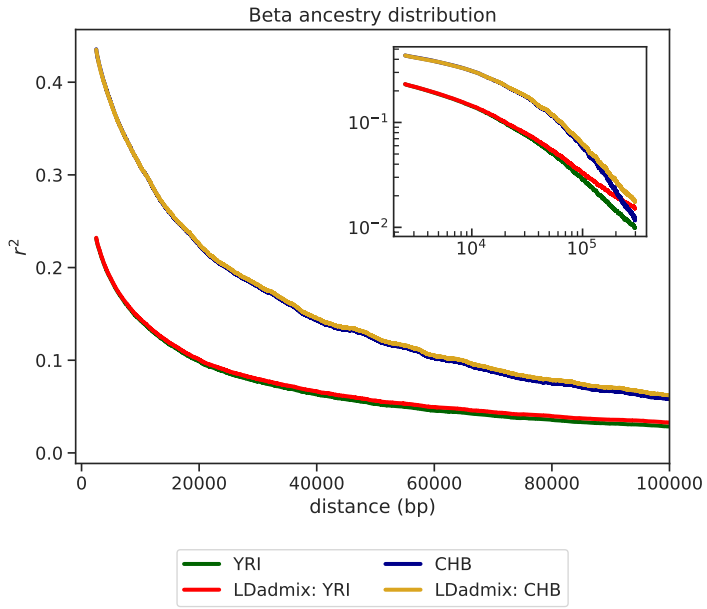
>0.05 and compare the resulting LD decay curves to those of the YRI and CHB source population samples to (Figure 3). Comparison to the LD decay curves estimated from YRI and CHB shows that LDadmix is able to recover the specific shape of the LD decay for both source populations, especially at distances less than 100kb (here equivalent to 0.1 cM), for both sets of Qs (Figure 3). At larger distances (100kb<x<300kb), with less LD in each source population, LDadmix also recovers the shape, but we see some upward bias in the estimates of r^2 (insets of Figure 3). Overall, our bias when estimating r^2 was positive, except in the highest LD bins of the CHB population, where we slightly underestimate true LD.

To quantify the accuracy of the estimates, we calculated RMSD and bias in distance bins of 5kb, or across all pairs as appropriate, using the LD from YRI and CHB as the truth. The two populations have some differences in their patterns of RMSD and bias. We see the population with less LD (YRI) has a more constant RMSD and mean bias across distance bins, while the population with more LD (CHB) has more variation in RMSD and bias (Figure 4). Also, while RMSD is consistently higher across all distance bins in the CHB populations, bias has a more complicated pattern that depends on distance.

Although the same haplotypes are used in both simulated scenarios there is an effect of the two different sets of admixture proportions, with overall RMSD and bias both higher in the uniform(0,1) scenario where the admixed individuals have more similar admixture proportionsRMSD 0.018 vs 0.036, bias 0.004 vs 0.009]. This effect is consistent across the range of distance bins examined (Figure 4), with the RMSD and bias of the beta ancestry distribution consistently lower for both populations across the range of distances. Across both analyses and both populations, RMSD is slightly smaller for loci at further distance (i.e. loci with lower mean r^2), but the bias is a bit higher (Figure 4).

(WAPLES ET AL., IN PREPARATION)

(A)



(B)

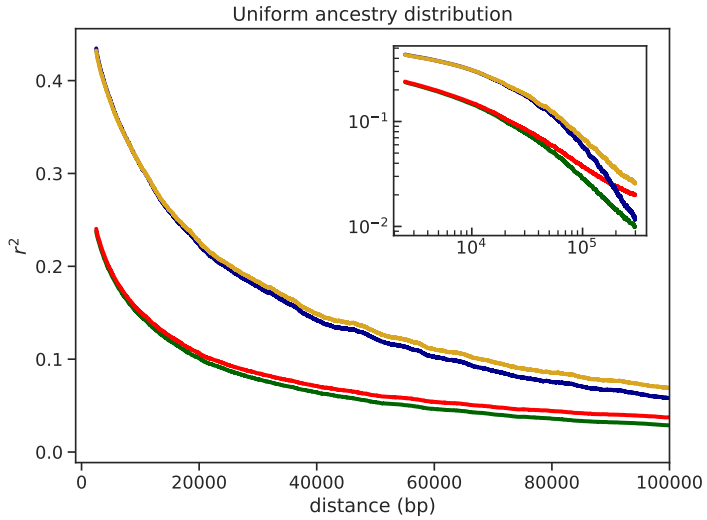


Figure 3: Estimated LD decay curves for both source populations in each of two simulated datasets. The red and yellow lines show mean estimated r^2 values in 5kb distance bins, the green and blue lines show the 'true' sample r^2 within each source population. The large plot shows the 5kb unweighted moving average of r^2, up to 100kb with a linear scale, while the inset plot extends the distance to 300kb and log-scales both the x- and y-axes. **A**) results for individuals simulated with a beta(0.1, 0.1) distribution of ancestry proportions, **B**) results for individuals simulated with a uniform(0,1) distribution of ancestry proportions.

(WAPLES ET AL., IN PREPARATION)

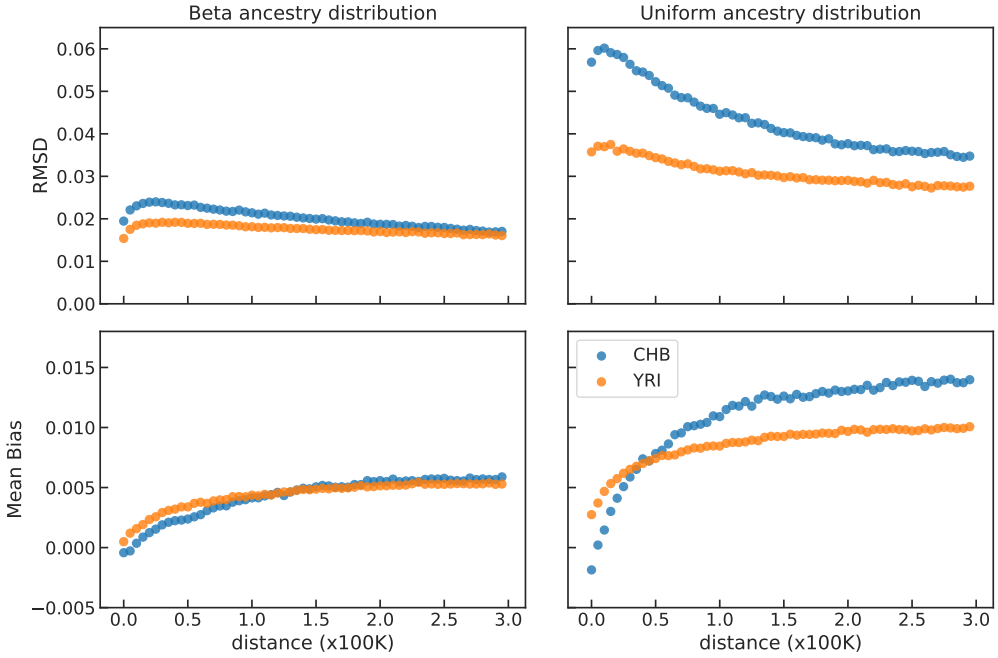


Figure 4: Scatterplots of RMSD, and mean bias in estimated r^2, in the two simulated scenarios. Orange points show deviations for the YRI, blue dots show the deviations for CHB, calculated in 5kb distance bins. The left column shows results for individuals simulated with a beta(0.1, 0.1) , the right column shows results for individuals simulated with a uniform(0,1) ancestry distribution.

To further investigate the effect of the ancestry distribution across a broader range of cases, we evaluated two additional simulation scenarios: 1) a low variance triangle ancestry distribution, with most individuals having close to the mean Q value of 0.5 (Supplemental Figure S2), and 2) supplementing all previous scenarios with eight additional unadmixed individuals from YRI, mimicking a case when additional samples from one source population are available.

Our ability to estimate r^2 when all individuals are near the mean ancestry proportion (triangle distribution) is significantly impacted, with consistently high RMSD and large mean bias across nearly all distance bins (Supplemental Figure S2). While were are able to identify the presence of a high LD and low LD population, the r^2 estimates for these two populations are biased towards each other at short distances, likely reflecting our inability to resolve their distinct haplotype frequencies. The asymptote of estimated r^2 for these two populations is also elevated, compared to the other sets of Qs. With the addition of eight additional unadmixed YRI individuals, we see a consistent reduction in RMSD in all scenarios and in both populations. The RMSD was reduced (0.182 vs 0.184) in the beta distribution, and was further reduced (0.033 vs 0.036) in the uniform ancestry distribution (Supplemental Figures S3, S4).

Application to 1000 Genomes data

To illustrate the utility of LDadmix, we applied it to admixed human populations from the Americas with the aim of recovering the ancestral pattern of LD in the Native American ancestry component while accounting for recent admixture with Africans and Europeans. In particular, we estimated LD decay in the Native American ancestry component of four populations from the 1000 Genomes project (PUR n=104, MXL n=61, PEL n=81, CLM n=94). To do so we first characterized the admixture in these populations with a K=3 unsupervised admixture analysis applied to samples from these American populations combined with three European and two African populations (Figure 5). The sum of the ancestry fractions for each ancestry are Native American: 138, African 233, and European 473. Notice that the majority

of the Native American ancestry component occurs in admixed individuals containing all three ancestry components, in contrast to the other ancestries. In fact, only 19 Native American individuals are estimated to be unadmixed, with more than 99% Native American ancestry: 18 from PEL and one from MXL. We then estimated the LD decay in each of the three ancestry components (American, African, European) shown in Figure 3A, by applying LDadmix to the genetic data as well as the estimated ancestry fractions (Figure 5).

To assess and interpret the results we first compared these curves to LD curves estimated directly from the African, European and Native American populations. Importantly, we find that the LD decay curves for the African and European ancestry components estimated by LDadmix coincides with the LD decay curves estimated from unadmixed samples (Figure 6A). This match is especially close for loci located close to each other, while at larger distances LDadmix estimates lower mean r^2 than we see in each individual population from each ancestry, likely due to the increased sample size in the combined analysis. We also see a similar pattern for the Native American ancestry if we compare it to the LD decay in the 18 unadmixed individuals from PEL, with a relatively close match at shorter distances that becomes worse at larger distance. However, here a direct comparison is more difficult because there are so few unadmixed Native American samples available and r^2 estimates are biased upwards with small sample sizes (Weir 1979). All in all, these observations support our simulation results in the sense that is suggest that LDadmix is providing meaningful LD decay curves.

Next we compared the LD decay curves obtained with LDadmix to LD decay measured within each of the admixed American populations alone using standard estimates of r^2 (Figure 6B). Among the admixed populations, the PEL population has the highest amount of LD, and also the highest fraction of Native American ancestry, and is closest to the LD estimated by LDadmix for Native American ancestry. The LD decay curves for the other American populations (MXL, CLM, PUR) all fall close to the r^2 values for the European ancestry, especially for loci at close distances, despite a having a very different demographic history from European. This may reflect the fact that these population samples have a substantial African and European admixture (Figure 5). At longer distances, these American populations have higher LD than Europeans, likely due to a combination of their history of admixture, smaller longer term N_e and reduced sample size in the data set. Notably, for all the admixed American populations, r^2 estimated with PLINK is below the r^2 estimated by LDadmix for the Native American ancestry across the entire range of distances, demonstrating that taking admixture into account makes a marked difference in the estimation of LD decay in these populations.

(WAPLES ET AL., IN PREPARATION)

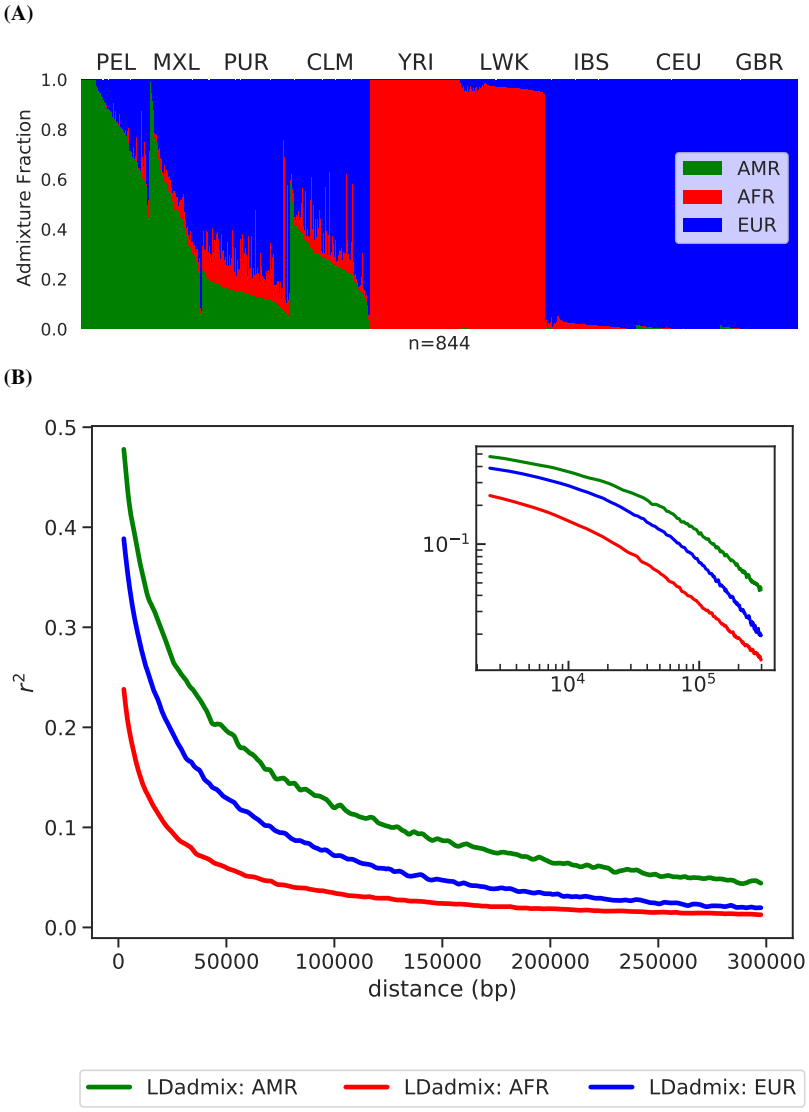


Figure 5: A) ADMIXTURE and LDadmix estimates for nine 1000 Genomes populations assuming three ancestral populations. Bar plot of a K=3 ADMIXTURE analysis of nine population samples from the 1000 Genomes. Green shading shows the estimated proportion of Native American ancestry, red shading shows the estimated proportion of African ancestry, blue shading shows the estimated proportion of European ancestry. Individuals are sorted by population sample (labels along the top) and then by the estimated proportion of Native American ancestry. **B**) LD decay curves estimated with LDadmix for the three ancestry components shown at the top, with matching colors. The large plot shows the 5kb unweighted moving average of r^2 with a linear scale, while the inset plot log-scales both the x- and y-axes.

(WAPLES ET AL., IN PREPARATION)

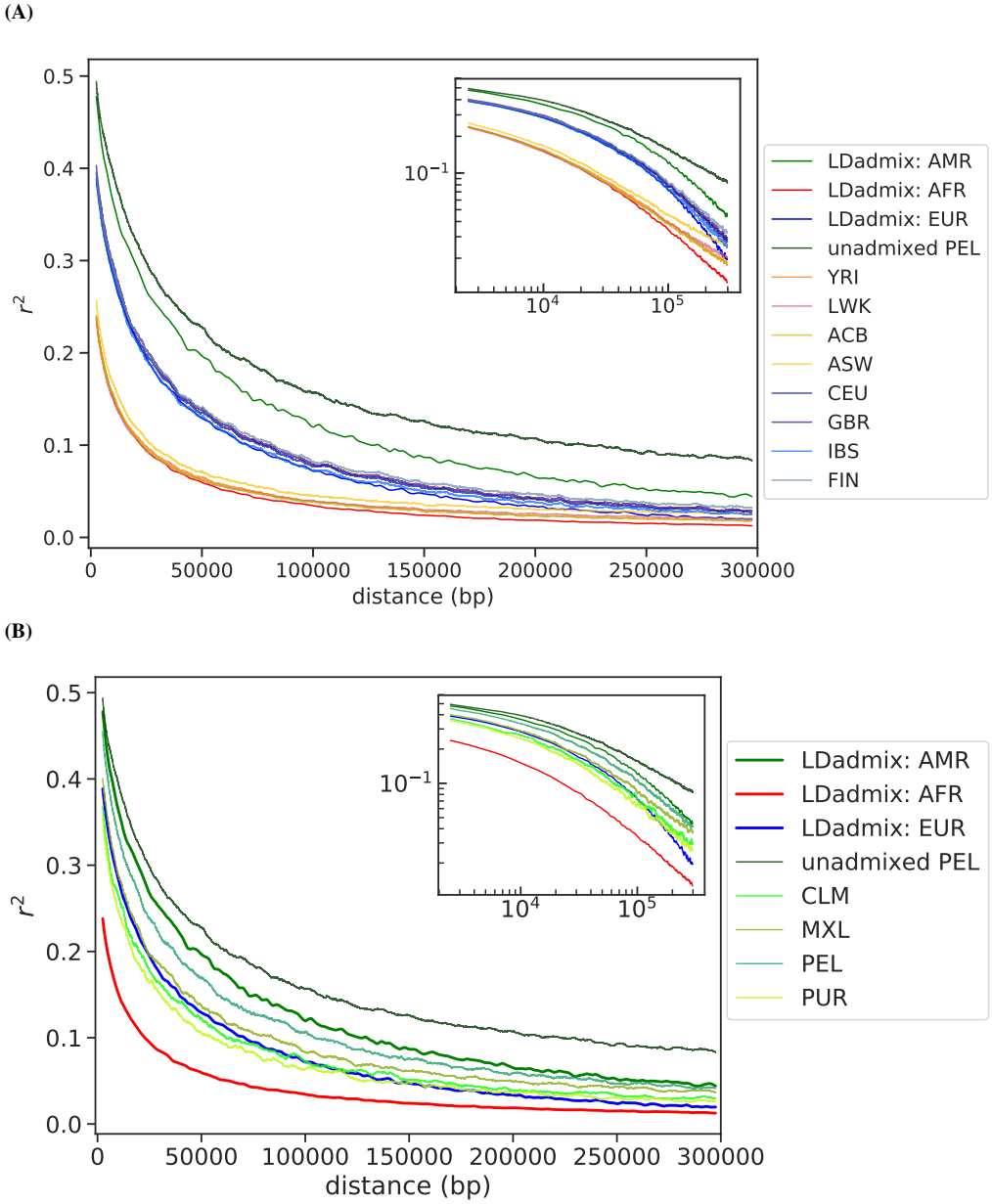


Figure 6: A) Standard LD estimates for nine 1000 Genomes population samples, plotted alongside LD decay curves estimated with LDadmix for the three ancestry components (Figure 5). The dark green line shows LD decay measured across 18 unadmixed individuals from PEL, representing an unadmixed sample from Native American ancestry. African-ancestry population samples are in red-yellow shades, European-ancestry populations are in blue-grey shades. The large plot shows the 5kb unweighted moving average of r^2 with a linear scale, while the inset plot log-scales both the x- and y-axes. **B**) LD decay for four American populations samples from the 1000 Genomes, plotted alongside LD decay curves estimated with LDadmix for the three ancestry components shown in Figure 5. The dark green line shows LD decay measured across 18 unadmixed individuals from PEL, representing an unadmixed sample from Native American ancestry. Other American populations samples are shown in shades of green. The large plot shows the 5kb unweighted moving average of r^2 with a linear scale, while the inset plot log-scales both the x- and y-axes.

(WAPLES ET AL., IN PREPARATION)

To test the robustness of the LD decay curve estimate obtained for the Native American ancestry component we also tried to apply LDadmix to the data without including the 18 unadmixed PEL individuals. This gives very similar results, suggesting that the LD estimates are not simply driven by the presence of these individuals (Supplemental Figure S5). We also applied LDadmix including only one American population at a time. The LD decay estimated for the Native American ancestry when the PEL individuals were the only included Americans was similar to the estimates obtained when the MXL, CLM and PUR individuals were also included, but the separate analysis of the other population samples provided less useful results (Supplemental Figure S6). These differences are likely due to the limited amount of Native American ancestry in the other population samples.

Software Availability

We have implemented LDadmix in a Python 3.X command line program. The program uses common file formats, and requires genotype data in binary PLINK format (link) as well as estimates of admixture proportions in the file format of ADMIXTURE (Q files). The program utilizes multiprocessing and is fast enough to be used on data sets with millions of pairs of loci. Shown in Table 1 are running times and peak memory usage using LDadmix to estimate LD at 1M pairs of loci in the 1000G K=3 dataset (n=844, L=500K), described in the manuscript. LDadmix is open source and available on github: https://github.com/rwaples/ancLD

Table 1: Running time and memory usage of LDadmix, as a function of number of cores on an example dataset. LDadmix was applied to a binary PLINK fileset with 844 individuals and 500K loci, and asked to estimate LD at 1M locus pairs.

Cores	Peak memory usage (GB)	Running Time (seconds)
1	8.3	4210
4	8.4	1172
16	8.5	394
64	8.4	200

Discussion

We have presented a software tool, LDadmix, that can provide estimates of LD in each of the ancestry sources of a recently admixed population. It requires only unphased genotype data and estimated admixture proportions for samples from the admixed population. Using simulated data, we have shown that the method can provide estimates of r^2 decay curves that are very close to those estimated from the true source populations for each ancestry in a K=2 admixture. Finally, by applying LDadmix to data from the 1000 Genomes project we obtained novel results for the Native American population ancestral to Peruvians, Mexicans, Colombians and Puerto Ricans as well as further insights into the benefits and the limitations to LDadmix.

Application to simulated data showed that the accuracy of the LD decays curves obtained with LDadmix depends on the distribution of admixture proportions. This makes sense, because compared to LD estimation methods that do not account for admixture, LDadmix has an additional challenge, as the source ancestry of each two-locus haplotype is not known, but rather the uncertainty is modeled probabilistically using the global ancestry fractions for each individual. In this model, a higher variance in admixture proportions values across individuals, provides more information about the source ancestry of each haplotype. Indeed we see that the presence of some essentially unadmixed individuals leads to more accurate LD estimates (Figure 2), and encouragingly, we also see that LDadmix is able to provide LD decay curves that are quite close to the truth, even when very few -if any- unadmixed samples are available (Figure 3B) as long as the samples do not all have similar ancestry proportions (Supplemental Figure S2). Furthermore, we see that including additional unadmixed samples from just one of the ancestral population seemed to improve LD estimates for both ancestral populations, suggesting a potential way improve results if one of the source populations is easy to get unadmixed samples from (Supplemental Figures S3, S4).

The simulation results also suggest that LDadmix performs a bit worse (higher absolute bias, higher RMSD) as loci get further apart and have less LD, for a few reasons. First, there is a greater chance of recombination between the loci since admixture, violating a key assumption of the model underlying LDadmix. Second, and more importantly, with less LD in the source populations, biases in LD estimation due to sample size contribute a larger fraction of the overall LD. Generally, when interpreting LD decay curves one should be cautious when using a low number of individuals. When using LDadmix to estimate LD for pairs of loci this issue is even more important since the effective number of individuals will be lower than the actual number of individuals, and dependent on the ancestry proportions. This should especially be taken into account when interpreting loci with very low amounts of LD or that are too far apart.

There are also several notable observations from applying LDadmix to admixed American population samples in the 1000 Genomes. First, the LD curves obtained with LDadmix for the African and European ancestries are very similar to those obtained directly from each of the African and European population samples (Figure 6A). These results show a consistent pattern of LD decay within the population samples from each continent and that LDadmix is able to recover this pattern. These results from a complicated K=3 admixture are likely significantly aided by access to many unadmixed European and African individuals, making it much easier to characterize the haplotype frequencies in these ancestries, and leaving the haplotype frequencies in the Native American ancestry as the principal unknown.

Second, although it is difficult to completely validate the LD decay curve obtained with LDadmix for the Native American ancestry due to the lack of unadmixed individuals, it does seem reasonable for several reasons. One reason is that this LD decay curve lies above all the LD curves obtained by estimating LD in these groups in the standard way without accounting for admixture (Figure 5). And importantly, the LD decay curves for these admixed American populations follow a pattern that those with the most European and African ancestry (PUR and CLM), have the lowest LD, closer to the LD patterns seen in European and African populations, while the population samples with more Native American ancestry (MXL and PEL) have more LD, in line with previous findings (Bryc et al. 2010). This pattern is consistent with the LD decay curves presented in Figure 1, with values in the admixed population that are intermediate compared to the source populations. Another reason is that, the LD decay curve obtained with LDadmix for the Native American ancestry is consistent with the LD decay curve obtained the 18 unadmixed PEL individuals,

although as previously explained this comparison is complicated by the small sample size. Finally, the elevated LD decay curve for the Native American ancestry estimated by LDadmix fits very well with current historical knowledge about the demographic history of Native American populations, including a bottleneck that is not shared by European, African, or Asian populations (Raghavan et al. 2015). Hence, all in all we believe the results are quite reasonable and this means that the results provide new insights into those populations, which clearly extends previous insights based on the admixed samples.

The results from the 1000 Genomes data also demonstrate the practical difficulties in generating reasonable estimates of LD decay in admixed populations without LDadmix. The LD within each American population is clearly affected by admixture, as seen above, but excluding the admixed individuals leaves a reduced sample size that can impact the results. For example, across the PUR, CLM, and MXL population samples there is only a single unadmixed individual, making LD estimation from only unadmixed individuals in these three populations essentially impossible. And even the PEL population sample, with 18 unadmixed individuals produced an LD decay curve with an elevated horizontal asymptote, a signature of limited sample size, complicating a comparison to LD curves in other populations produced with larger sample sizes.

Finally, the 1000 Genomes data analysis reveal some practical limitations to keep in mind when applying LDadmix. The estimate of LD decay for the Native American ancestry derived from the pooling of all the Native American population samples (Figure 5) was distinct from the results when applying LD to each American population separately (Supplemental Figure S6). These differences in LD may reflect the distinct demographic history of each population, however, the limited Native American ancestry within each population sample presents two further challenges to LD estimation, both related to effective sample size, that we think are likely to have a larger effect. First, due to the sensitivity of LD estimators to sample size, LD comparisons between populations are often matched in sample size (see e.g. 1000 genomes paper) and this matching is especially important with lower sample sizes. However in admixed populations, matching sample sizes becomes much more difficult, because the number of haplotypes with a certain ancestry will vary along the genome. Second, in LDadmix the source ancestry of each two-locus haplotype is not known, it is modelled probabilistically using the global ancestry fraction and therefore the LDadmix estimates of haplotype frequencies and LD are associated with additional uncertainty, reducing the effective sample size of each ancestry. This results in a further upwards bias in LD, visually evident as the horizontal asymptote in the estimated r^2 values, especially visible on the inset log-scaled plots (Figure 6, Supplemental Figures S2,S3,S6). Several methods exist that attempt to correct for the effect of sample size on the estimation of population r^2 (Weir and Hill 1980; Waples 2006; Bulik-Sullivan et al. 2015; Ragsdale and Gravel 2019), however, it is difficult to apply downsampling and correction methods to r^2 estimates from LDadmix because there is not a single sample size across loci. It may be possible to use the r^2 value at the asymptote as a part of a future sample size correction scheme in LDadmix, however, we have not yet managed to successfully develop such a scheme. We therefore suggest only to use LDadmix on large samples sizes, like we did in the case of the dataset of all Native American population samples from the 1000 genomes project.

Despite this limitation we believe these results show LDadmix has the potential to be a useful tool for the analysis of LD in admixed populations. Current studies of LD in admixed populations face a difficult choice: to either include admixed individuals or to exclude all admixed individuals. But if admixed individuals are included in LD calculations, the LD becomes difficult to interpret as it reflects both the demographic and admixture history of the individuals and if admixed individuals are excluded, the available size maybe too small for useful measures of LD. LDadmix presents a solution to this problem, by accounting for admixture in the estimation of the LD it presents the ability to better utilize all available individuals, expanding sample sizes and improving LD estimates. Therefore, we believe it has potential to be a useful tool in future studies – both of humans and especially in other species where large panels of unadmixed individuals are not available.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." Nature 526 (7571): 68–74.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." Genome Research 19 (9): 1655–64.

Bryc, Katarzyna, Christopher Velez, Tatiana Karafet, Andres Moreno-Estrada, Andy Reynolds, Adam Auton, Michael Hammer, Carlos D. Bustamante, and Harry Ostrer. 2010. "Genome-Wide Patterns of Population Structure and Admixture Among Hispanic/Latino Populations." In In the Light of Evolution: Volume IV: The Human Condition, edited by John C. Avise and Francisco J. Ayala. Washington, DC: The National Academies Press.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies." Nature Genetics 47 (3): 291–95.

Chakraborty, R., and K. M. Weiss. 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 of the United States of America 85 (23): 9119–23.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." GigaScience 4 (February): 7.

Excoffier, L., and M. Slatkin. 1995. "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population." Molecular Biology and Evolution 12 (5): 921–27.

Gaunt, Tom R., Santiago Rodríguez, and Ian Nm Day. 2007. "Cubic Exact Solutions for the Estimation of Pairwise Haplotype Frequencies: Implications for Linkage Disequilibrium Analyses and a Web Tool 'CubeX.'" BMC Bioinformatics 8 (November): 428.

Grey, Corinne, Frédéric Baudat, and Bernard de Massy. 2018. "PRDM9, a Driver of the Genetic Map." PLoS Genetics 14 (8): e1007479.

Gutenkunst, Ryan N., Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. 2009. "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data." PLoS Genetics 5 (10): e1000695.

Hellenthal, Garrett, George B. J. Busby, Gavin Band, James F. Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. 2014. "A Genetic Atlas of Human Admixture History." Science 343 (6172): 747–51.

Hill, William G. 1981. "Estimation of Effective Population Size from Data on Linkage Disequilibrium." Genetics Research 38 (3): 209–16.

Kawakami, Takeshi, Niclas Backström, Reto Burri, Arild Husby, Pall Olason, Amber M. Rice, Murielle Ålund, Anna Qvarnström, and Hans Ellegren. 2014. "Estimation of Linkage Disequilibrium and Interspecific Gene Flow in Ficedula Flycatchers by a Newly Developed 50k Single-Nucleotide Polymorphism Array." Molecular Ecology Resources 14 (6): 1248–60.

Kelleher, Jerome, Alison M. Etheridge, and Gilean McVean. 2016. "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes." PLoS Computational Biology 12 (5): e1004842.

Loh, Po-Ru, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K. Pickrell, David Reich, and Bonnie Berger. 2013. "Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium." Genetics 193 (4): 1233–54.

Maples, Brian K., Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. 2013. "RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference." American Journal of Human Genetics 93 (2): 278–88.

McVean, Gilean A. T. 2002. "A Genealogical Interpretation of Linkage Disequilibrium." Genetics 162 (2): 987–91.

Moltke, Ida, Matteo Fumagalli, Thorfinn S. Korneliussen, Jacob E. Crawford, Peter Bjerregaard, Marit E. Jørgensen, Niels Grarup, et al. 2015. "Uncovering the Genetic History of the Present-Day Greenlandic Population." American Journal of Human Genetics 96 (1): 54–69.

Moorjani, Priya, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po-Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh. 2013. "Genetic Evidence for Recent Population Mixture in India." American Journal of Human Genetics 93 (3): 422–38.

Myers, Simon, Charles Fefferman, and Nick Patterson. 2008. "Can One Learn History from the Allelic Spectrum?" Theoretical Population Biology 73 (3): 342–48.

Plagnol, Vincent, and Jeffrey D. Wall. 2006. "Possible Ancestral Structure in Human Populations." PLoS Genetics 2 (7): e105.

Pritchard, J. K., and M. Przeworski. 2001. "Linkage Disequilibrium in Humans: Models and Data." American Journal of Human Genetics 69 (1): 1–14.

Raghavan, Maanasa, Matthias Steinrücken, Kelley Harris, Stephan Schiffels, Simon Rasmussen, Michael DeGiorgio, Anders Albrechtsen, et al. 2015. "Genomic Evidence for the Pleistocene and Recent Population History of Native Americans." Science 349 (6250): aab3884.

Ragsdale, Aaron P., and Simon Gravel. 2018. "Models of Archaic Admixture and Recent History from Two-Locus Statistics." bioRxiv . https://doi.org/10.1101/489401.

Ragsdale, Aaron P., and Simon Gravel. 2019. "Unbiased Estimation of Linkage Disequilibrium from Unphased Data."bioRxiv. https://doi.org/10.1101/557488.

Ragsdale, Aaron P., and Ryan N. Gutenkunst. 2017. "Inferring Demographic History Using Two-Locus Statistics." Genetics 206 (2): 1037–48.

Slatkin, Montgomery. 2008. "Linkage Disequilibrium--Understanding the Evolutionary Past and Mapping the Medical Future." Nature Reviews. Genetics 9 (6): 477–85.

Supple, Megan A., and Beth Shapiro. 2018. "Conservation of Biodiversity in the Genomics Era." Genome Biology 19 (1): 131.

Tenesa, Albert, Pau Navarro, Ben J. Hayes, David L. Duffy, Geraldine M. Clarke, Mike E. Goddard, and Peter M. Visscher. 2007. "Recent Human Effective Population Size Estimated from Linkage Disequilibrium." Genome Research 17 (4): 520–26.

Waples, Robin S. 2006. "A Bias Correction for Estimates of Effective Population Size Based on Linkage Disequilibrium at Unlinked Gene Loci*." Conservation Genetics 7 (2): 167.

Waples, Robin S., and Chi Do. 2010. "Linkage Disequilibrium Estimates of Contemporary N E Using Highly Variable Genetic Markers: A Largely Untapped Resource for Applied Conservation and Evolution." Evolutionary Applications 3 (3): 244–62.

Weir, B. S. 1979. "Inferences about Linkage Disequilibrium." Biometrics 35 (1): 235–54.

Weir, B. S., and W. G. Hill. 1980. "Effect of Mating Structure on Variation in Linkage Disequilibrium." Genetics 95 (2): 477–88.

(WAPLES ET AL., IN PREPARATION)

Supplemental Figures

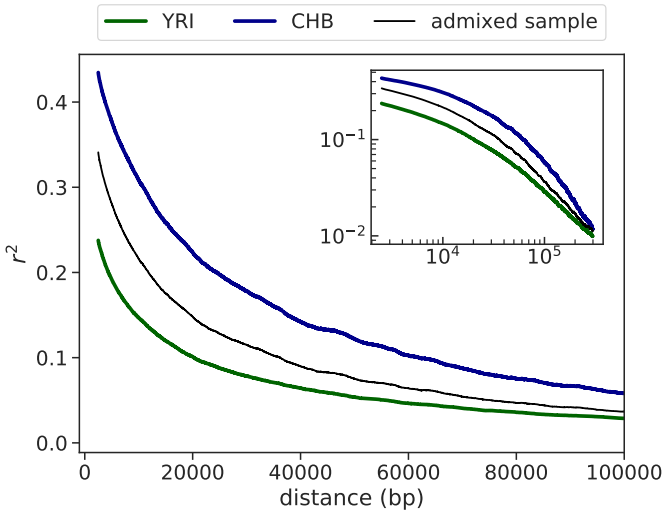


Figure S1: LD decay curves for source (blue and green, N=100) and admixed (admixed, n=200) populations in simulation based analysis. The large plot shows the 5kb unweighted moving average of r^2, up to 100kb with a linear scale, while the inset plot extends the distance to 300kb and log-scales both the x- and y-axes.

(WAPLES ET AL., IN PREPARATION)

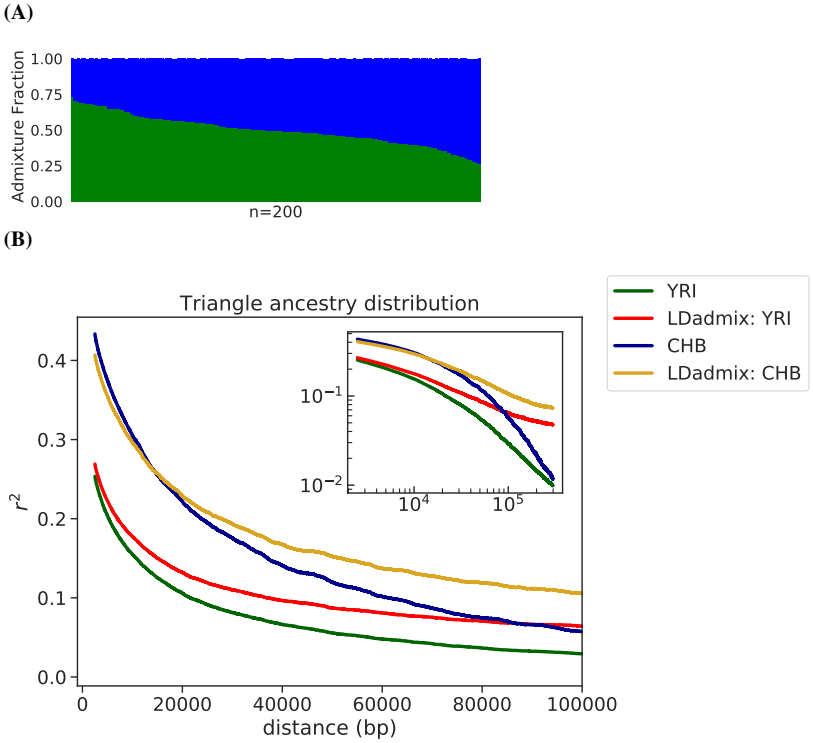


Figure S2: LDadmix applied to simulated data with the triangle (mode = 0.5, min = 0.25, max = 0.75) admixture proportion distribution. A) Bar plot of the admixture proportions in this scenario, showing the admixture proportions for 200 individuals. B) Estimated LD decay curves for both source populations. The red and yellow lines show estimated r^2 values in 5kb bins, the green and blue lines show the 'true' sample r^2 within each source population, as in figure 2B. The large plot shows the 5kb unweighted moving average of r^2, up to 100kb with a linear scale, while the inset plot extends the distance to 300kb and log-scales both the x- and y-axes.

(WAPLES ET AL., IN PREPARATION)

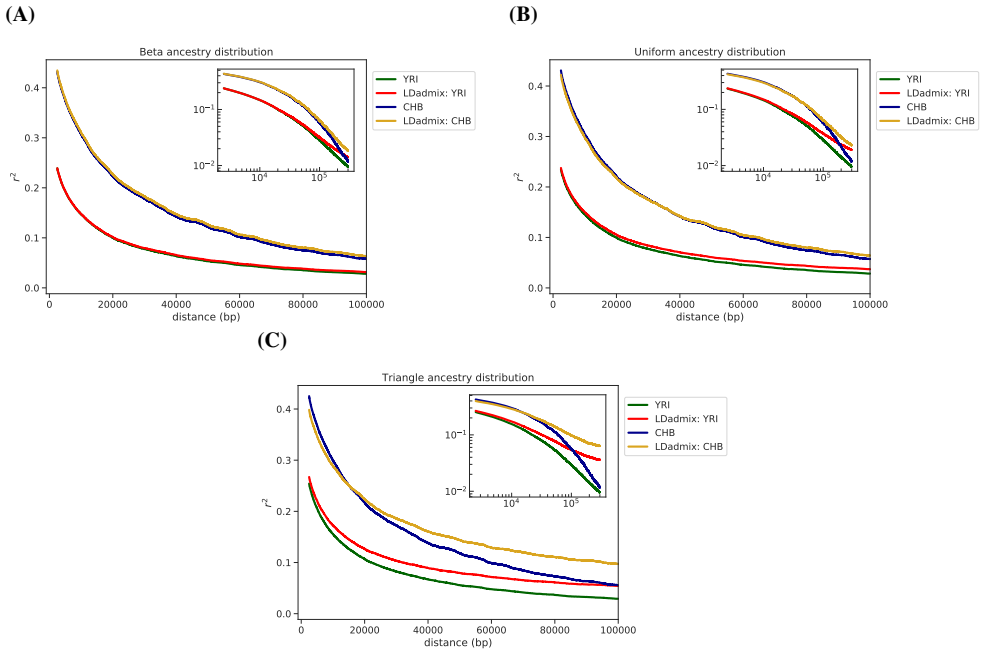


Figure S3: Effect of adding 8 additional unadmixed YRI individuals. Estimated LD decay curves for both source populations. The red and yellow lines show mean estimated r^2 values in 5kb bins, the green and blue lines show the 'true' sample r^2 within each source population, as in Figure 3. The large plot shows the 5kb unweighted moving average of r^2, up to 100kb with a linear scale, while the inset plot extends the distance to 300kb and log-scales both the x- and y-axes. **a**) Admixture proportions for 200 individuals drawn from a beta(0.1, 0.1), **b**) Admixture proportions for 200 individuals drawn from a uniform(0,1), **c**) Admixture proportions for 200 individuals drawn from a triangle distribution with mode 0.5, min = 0.25, max = 0.75. Compare to figure 3A, B in the main text.

(WAPLES ET AL., IN PREPARATION)

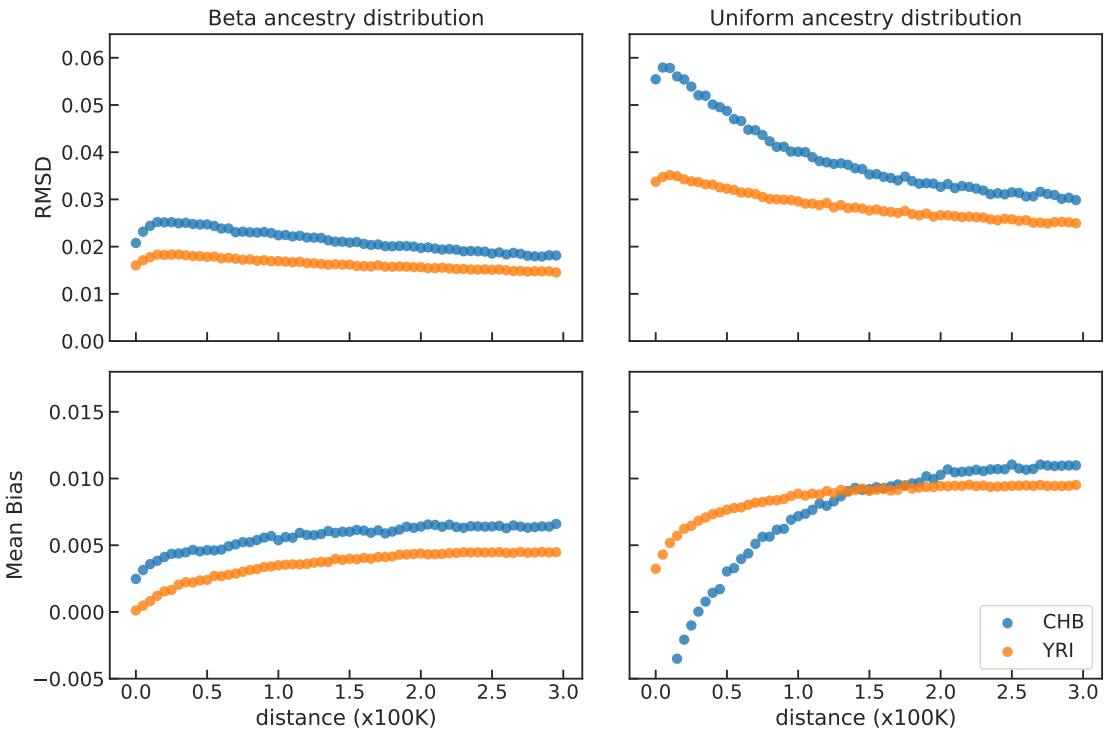


Figure S4: Effect of adding 8 additional YRI individuals. Scatterplots of RMSD and mean bias in estimated r^2 in the two simulated scenarios. Orange points show deviations for the YRI, blue dots show the deviations for CHB, calculated in 5kb distance bins. The left column shows results for individuals simulated with a beta(0.1, 0.1) , the right column shows results for individuals simulated with a uniform(0,1) ancestry distribution. Compare to Figure 4 in the main text.

(WAPLES ET AL., IN PREPARATION)

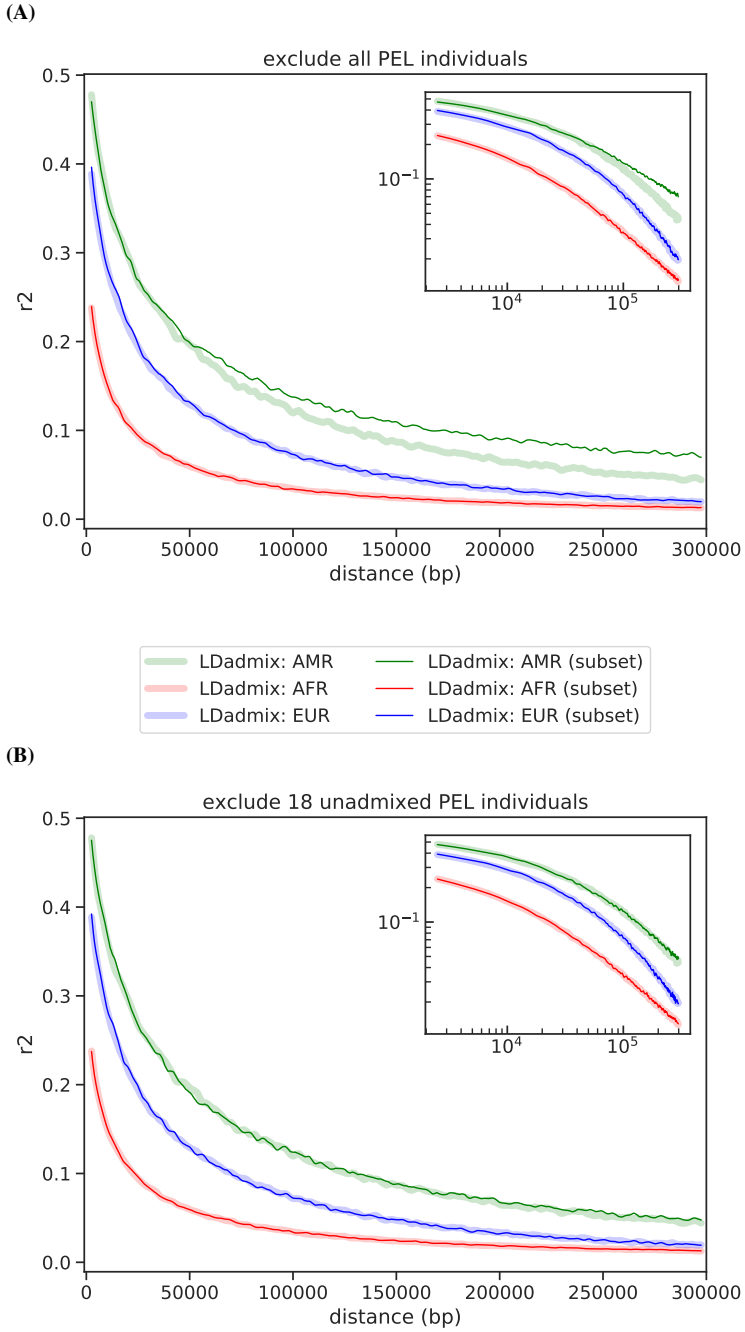


Figure S5: Effect of leaving out two sets of PEL individuals on LD decay estimates. **a**) excluding all PEL individuals, **b**) excluding the 18 unadmixed PEL individuals with >99% Native American ancestry. Faint lines show the LD decay curves estimated with LDadmix on the entire data set, matching those shown in Figure 5. Thin bright lines show the estimates for the particular data set show on each plot. The large plot shows the 5kb unweighted moving average of r^2 with a linear scale, while the inset plot log-scales both the x- and y-axes.

(WAPLES ET AL., IN PREPARATION)

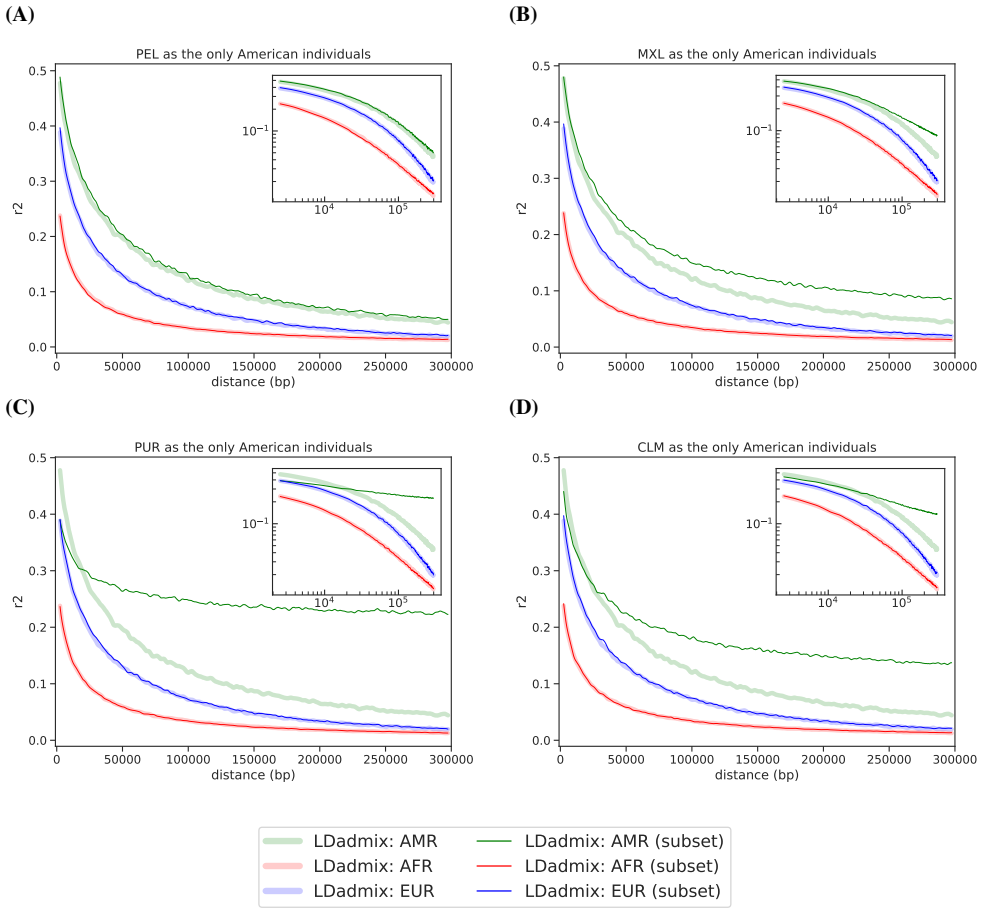


Figure S6: Effect of leaving out all individuals from each American population sample. Faint lines show the LD decay curves estimated with LDadmix on the entire data set, matching those shown in Figure 5. Thin bright lines show the estimates for the particular data set show on each plot. The large plot shows the 5kb unweighted moving average of r^2 with a linear scale, while the inset plot log-scales both the x- and y-axes. **a**) PEL individuals were the only American individuals included. **b**) MXL individuals were the only American population individuals included **c**) PUR individuals were the only American individuals included **d**) CLM individuals were the only American individuals included.

Acknowledgements

There are many people that played a notable part in these three years, and they all deserve some credit.

First, I would like to thank my supervisor Ida for always having the time, patience, and rigor to be a proper supervisor, in all aspects of science. And thanks also to Anders, for stocking the candy store, to cutting through to the important stuff (mostly), and always being ready to throw a quick party. I benefited greatly from working with both of you.

It was a delight to work with all my fellow grad students and office mates - Emil, for help with merging and always having a new word to discuss - Casper-Emil for sharing beers and advice, Samuele for always having a (pleasant) opinion, Jonas, for setting a good example in Python - Yorgos for support and advice and care, Lys for keeping the mood in the office a bit lighter, Patrícia for bringing the board games, and all the rest that keep that big office an enjoyable place to work - whether you are actually in the office or not - Christian, Ninna, Anne, Genis, Mengyuan.

All the rest of the section at KU, you make me feel welcome here - Henrietta for all the little things - Hans, for looking out for me, and Rasmus and Rute for setting a good example in science and in humor. I want to thank Fernando for hosting the best journal club in Copenhagen and Aviaja for being a pleasure to work with on the Greenlandic project.

I would also like to express my gratitude to all the people at UCL that made my stay there an enjoyable and productive one. Garrett, thank you for hosting me and for the engaging discussions. Lucy, I think you saved me at least two months of work with your sage advice and experience. Agata, for making my stay a breeze, and Javier, Dave and the rest for making it a pleasure.

Moving to a new country is not always so easy, and nobody knew what I was going through quite like the rest of my Introduction to Danish class. Rob, Montse, and Patrica, especially. I also need to thank team EasyOn - all the running I have to do to make up for my football skills keeps me sane and almost fit.

Morten and Maria and Balder - thank you for showing me my first slice of Denmark, all the way back in Seattle, and for being my home away from home. I haven't seen you as much as I thought I would her in Denmark, and that's too bad.

I owe a lot to my early scientific mentors, I wouldn't be here without you - Jim and Lisa, Linda, and Sonia, thank you. That includes you too dad.

Tone, for showing a remarkable amount of care, no matter what.

And finally to my family - to my parents, for their constant support, and to my sister for putting up with me.

Appendix A:
Two additional papers co-authored during the PhD

Conservation Genetics (2019) 20:101–113
https://doi.org/10.1007/s10592-019-01149-5

RESEARCH ARTICLE

More grist for the mill? Species delimitation in the genomic era and its implications for conservation

David W. G. Stanton[1,2] · Peter Frandsen[3,4] · Ryan K. Waples[3] · Rasmus Heller[3] · Isa-Rita M. Russo[1] · Pablo A. Orozco-terWengel[1] · Casper-Emil Tingskov Pedersen[3] · Hans R. Siegismund[3] · Michael W. Bruford[1,5]

Received: 14 August 2018 / Accepted: 18 January 2019 / Published online: 1 March 2019

Abstract
Species delimitation is one of the most contested areas in modern biology, with widespread disagreement about almost every aspect of the definition and implementation of the "species" label. While this debate is intellectually stimulating, it also has real implications for conservation, where its impacts on taxonomic inflation or inertia can mean that specific populations receive adequate conservation measures or are ignored. Recently, the rise of next generation sequencing and phylogenomics has revolutionised phylogenetic understanding of many organismal groups but has simultaneously highlighted the porosity of genomes in terms of admixture across previously delineated species barriers. The extraordinary power of genomic data is increasingly being used to delineate species, and several publications in this domain have recently attracted significant attention and criticism. Here we revisit the question of species delimitation, but from a genomic context. We ask how and whether the large amounts of data provided by genomic methods can resolve the longstanding discussion on the validity and application of phylogenetic and allied species concepts, and how some recent examples can inform this debate. We argue that conserving adaptive potential is a priority for conservation, and no single species concept currently does that adequately on its own. Genomic data holds the potential to add unprecedented detail, but frequently falls short of this potential.

Keywords Genomics · Biological species concept · Phylogenetic species concept · Adaptive introgression · Hybridization

Inflation or inertia?

Due to the pivotal role of the species as the most important unit of biodiversity, conservation planning must be based on a good understanding of species number, diversity and endemism, measurements that only make sense within the context of consistent taxonomic classifications (Isaac et al. 2004; Zachos et al. 2013). However, as a result of the many different epistemological views on the species concept (e.g. Avise and Ball 1990; Wheeler and Platnick 2000; Baker and Bradley 2006), and due to the gradual process of evolutionary divergence, there is a continuum of genomic divergence patterns and estimates for which different researchers would consider speciation to be 'complete' (DeQueiroz 1998). Some evolutionary biologists have classified populations as the same species unless strong evidence to the contrary exists, *e.g.* reproductive incompatibility or reciprocal monophyly (with the archaic term 'lumpers'; Heller et al. 2013). The genealogical concordance method of phylogenetic species recognition (often known as the genealogical species concept, or GSC; Avise and Ball 1990; Baum and Shaw

Peter Frandsen and David W. G. Stanton contributed equally to this article.

✉ David W. G. Stanton
david.stanton@nrm.se; dave.stanton84@gmail.com

✉ Hans R. Siegismund
hsiegismund@bio.ku.dk

✉ Michael W. Bruford
brufordmw@cardiff.ac.uk

[1] Cardiff School of Biosciences, Sir Martin Evans Building, Cardiff University, Museum Avenue, Cardiff CF10 3AX, UK

[2] Department of Genetics and Bioinformatics, Swedish Museum of Natural History, Frescativägen 40, 114 18 Stockholm, Sweden

[3] Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

[4] Research and Conservation, Copenhagen Zoo, 2000 Frederiksberg, Denmark

[5] Sustainable Places Institute, Cardiff University, Cardiff CF10 3BA, UK

1995), as well as the Biological species concept (BSC), often result in a high threshold of species recognition. The GSC usually considers two populations to be designated species only when they are "isolated long enough [that] all gene-genealogies will be concordant" (emphasis ours; Baum and Shaw 1995). This concept has been criticized for its stringent nature, as it, for example conflicts with the observed incomplete lineage sorting and admixture between the genomes of some well-recognised species (for example lineage sorting in Ursine bears, Kutschera et al. 2014; and apes; Mailund et al. 2014; and introgression between chimpanzees and bonobos deManuel et al. 2016; within gulls; Sonsthagen et al. 2016; and geese; Ottenburghs et al. 2017).

In contrast, other evolutionary biologists set the threshold for recognition of new species, much lower (i.e. so-called 'splitters', the past antonym of 'lumpers') whose approach is usually via the use of the Phylogenetic Species Concept (PSC). The PSC defines species as "the smallest aggregation of (sexual) populations or (asexual) lineages diagnosable by a unique combination of character states" (Wheeler and Platnick 2000). This method of classification is much less stringent and it could be argued that any intra-specific population genetic structure should result in the fulfilment of the requirement of "a unique combination of character states". It has therefore been criticized for increasing the number of recognized species beyond what would seem justified, known as 'taxonomic inflation' (Heller et al. 2013; Zachos 2013; Zachos et al. 2013).

Recently, Gippoliti et al. (2017) describe the opposing argument that 'taxonomic inertia' is actually more detrimental to conservation, highlighting the case of African ungulates. They argue that the history of African ungulate taxonomic classification has been dominated by 'lumpers' who, when faced with difficult taxonomic decisions, have avoided the situation by assigning a large number of subspecies or genetic lineage labels. According to the authors, this has led to a disproportionately small number of ungulate species being recognised in Africa [despite Africa being by far the leading continent in terms of recognized ungulate species richness (Heywood 2010)]. Key to the argument of Gippoliti et al. (2017) is a survey by Morrison et al. (2009), which showed that taxonomic splitting has a positive effect on conservation. Morrison et al. (2009) identify numerous situations where a change in taxonomy has led to increased conservation efforts. One representative example is the California gnatcatcher, *Polioptila californica* Morrison et al. (2009) highlight the increase in conservation funding (better habitat protection and monitoring programs) that this species received after recognition of its species status. However, a change in protection (conservation) in this study was defined in Morrison et al. (2009) as "increased or decreased monitoring of any kind", as well as "increased or decreased funding for research on the respective organism". This argument only considers the organism in question, not conservation actions as a whole. This overlooks an obvious concern, namely that conservation resources are finite (although not necessarily constant), and that resources allocated to one species do not necessarily benefit others. This is the 'Agony of choice' argument (Isaac et al. 2004), which refers to the greater challenge of assigning limited conservation resources between higher numbers of taxa. Gippoliti et al. (2017) also state that there is "no evidence for negative effects of taxonomic splitting on conservation". It could be argued, however, that this hypothesis would be very difficult to empirically support either way. It is not currently known precisely how much is being spent on conservation globally (McCarthy et al. 2012), let alone the relative amounts that are being spent on each taxonomic group. In an ideal scenario, all units of diversity would be conserved however, even in countries that allocate a relatively large budget to conservation efforts, this is rarely possible (Malaney and Cook 2013), and even a prioritization approach may not be being practised (especially when it conflicts with other political priorities migration, denver post). It therefore seems very likely that conserving the eleven species of klipspringer proposed by Groves and Grubb (2011) as separate taxa would require more resources than conserving the one klipspringer species, *Oreotragus oreotragus*, commonly recognised (e.g. Kingdon 2013). In conservation (specifically in the IUCN context), a particular machinery comes into play when a new species becomes known, including making species status assessments, a species survival plan including in situ and ex situ measures (if deemed necessary for the species). All of these obviously require resources, and this is before even expending resources on the actual, practical conservation measures for the species.

Another argument for why over-splitting may be detrimental for particular taxonomic groups, Frankham et al. (2012) focused on three widely used species concepts: the Biological (Mayr 1942, 1963), the Evolutionary (ESC; Simpson 1951, 1961; Wiley 1978) and Phylogenetic (Eldredge and Cracraft 1980; Cracraft 1997) Species Concepts. Frankham et al. (2012) emphasised the point that diagnosably different population units are not intrinsically reproductively isolated (the importance of reproductive isolation is discussed in detail later), and that this is particularly relevant for small, allopatric populations. This is because the time in which a population attains alleles that make it distinguishable in e.g. multivariate genetic space will be proportional to their effective population sizes (Ne), and may be very short if Ne is low. The implication of this is that populations of the greatest conservation concern may be more likely to be diagnosed by the PSC. It should be noted that this argument is only referring to the conservation implications of the species concept used, and not an assessment of which is "correct".

It seems clear that both "taxonomic inflation" and "taxonomic inertia" could be detrimental to conservation. This is because implicit in those descriptors is an assumption that the populations in question have been artificially "grouped" more or less than what would be ideal under any given criterion (also keeping in mind that different criteria might not lead to the same "ideal" grouping!). Conservation (and in particular its genetic component) is predominantly a pragmatic discipline, which for populations threatened with extinction, a primary concern is assessing whether re-joining populations (and therefore gene flow) is possible and deciding whether those populations *should* still be connected. These decisions are usually based on taxonomy, with the assumption that it is related to whether the populations are likely to be reproductively isolated, and to what extent they have unique adaptations to their local environment. However, this assumption may be correct or incorrect, depending on the premise of the species criterion used (see "Which species concept best conserves adaptive potential?"). This raises three important distinctions that need to be made when a decision is reached about what constitutes a species:

1. Distinguishing species
2. How they are diagnosed
3. Classification, i.e. how they are ranked

Point one is an ontological question, i.e. what one considers a species to actually *be*. Point two is a question of implementation: a technical/financial hurdle that is imposed based on the species concept that is chosen. Point three could be referred to as a "convention of organization", and depends on where any given organization chooses to delimit taxonomic boundaries. This framework links to the difference between a concept and a criterion, two terms that are frequently conflated in species discussions: a species concept relates to point one, and a species criterion relates to point two (De Queiroz 1998).

Some authors may argue that certain criteria are invalid because they do not identify units that they believe to be "real" species, however this can be countered by defining criteria as a concept, and thereby essentially redefining what a species *is* to fit in with a given criterion. It has been argued that all species concepts have a single common concept, namely that species can be equated with "segments of population-level lineages" (De Queiroz 1998, 1999), or groups of organisms with their own "independent evolutionary fate and historical tendencies" (Mayden 1997). Under this framework, the so-called general lineage concept (GLC), it is argued that alternative species concepts are either variations of the GLC, or criteria of it. While this is a compelling argument, in the sense that it relates to ontology, it could still be considered a matter of opinion.

In an attempt to make the definition of species less arbitrary, increasingly sophisticated methods have been produced to delineate species. Developments in coalescent theory has allowed for the investigation of lineage diversification (Yang 2015). Other methods for molecular species identification include Bayes factor-based species delimitation (Grummer et al. 2014). These methods, based on different criteria/theories, test species boundaries in a comparative way (Toussaint et al. 2016). However, the multispecies coalescent has also been criticised for only being a method to "delimit structure, not species" (Sukumaran and Knowles 2017).

Further discussion on which of the various species concepts is "correct" remains outside the scope of this manuscript. Rather we seek to ask if, and how, genomic data have influenced the operational nature of the various species concepts. Specifically, has the increasing resolving power of genomic tools (i) been used to invoke the chosen species concept (we focus on the PSC and BSC) more readily, or, (ii) led to a more conservative approach to species delineation due to the complex interaction of admixture, incomplete lineage sorting, and demographic history that is increasingly being detected. We also revisit the question of the link between species concepts and adaptive potential, and whether new genomic data has had an influence on this question. We focus on case studies from the recent literature (Table 1), which highlight how species delineations have been applied to date. These studies either use what could broadly be described as the BSC or the PSC (here synonymous with "lumping" and "splitting", respectively) in order to justify their species delineations.

Newly proposed species

Giraffe

Recently, Fennessey et al. (2016), produced a draft genome for the giraffe (*Giraffa camelopardelis*), and analysed nuclear and mitochondrial sequences from 105 individuals from all currently recognized subspecies. In identifying four distinct genetic clusters they concluded that "population genetic, phylogenetic, and network analyses of nuclear sequences demonstrate that the giraffe is genetically well structured into four distinct species". However, this conclusion was based on only two mitochondrial and seven intron loci. It contrasts with a previous genetic study of giraffe, which used 14 microsatellite loci from 381 individuals to identify six distinct clusters (Brown et al. 2007), without designating these clusters to species. Therefore, both studies were based on a relatively small number of loci that showed varying genetic structure but reached different conclusions. This could be explained by variation among loci

Table 1 Summary of the genomic evidence used in our case studies

Study	Reference	Genomic resources	Analyses	Species criteria used
Newly proposed species				
Giraffe	Fennessy et al. (2016)	One draft genome	• Phylogeny • Genetic structure (Using 7 intron loci and mitochondrial DNA)	Unique genetic character states (PSC)
Orang-utan	Nater et al. (2017)	37 resequenced genomes	• Phylogeny • Genetic structure • Demographic history • Morphology (Genetic data from two, and morphological data from one *Pongo tapanulienis* individual[s])	Unique genetic and morphological character states (PSC)
Finless porpoise	Zhou et al. (2018)	48 resequenced genomes	• Phylogeny • Genetic structure • Demographic history • Signatures of selection	Reproductive isolation (BSC)
Darwin's finch spp.	Lamichhaney et al. (2017)	47 resequenced genomes, genomic data from 180 individuals from previous study	• Phylogeny • Morphology • Pedigree assessments • Demographic history • Phenotype-genotype associations	Reproductive isolation (BSC)
Currently recognised species				
Stickleback spp.	Ravinet et al. (2018)	27 resequenced genomes	• Phylogeny • Demographic history • Detection of introgression • Genetic structure • Signatures of selection	Species claim not made in this study, but well-recognised as different species with reproductive isolation, and ecological and phenotypic differences (BSC)

with different realisations of stochastic lineage sorting, an effect that while still possible for large numbers of loci, is more likely to be observed in studies using relatively few (Orozco-terWengel et al. 2011). The operational approach used in Fennessey et al. (2016) could be described as conforming to the PSC, as the genetic structure was used to justify a "unique combination of character states" (i.e. nuclear alleles), present in each of the populations (or species).

Of all the examples presented below, the findings presented in Fennessey et al. (2016) have probably received the most attention to date, reviving the debate on giraffe taxonomy and conservation. Bercovitch et al. (2017) listed seven points of concern about the original authors' interpretation of their results. Their concerns included a criticism of the lack of concordance between mitochondrial and nuclear phylogenies, few loci, and a disagreement that assignment to separate genetic clusters was a sufficient indicator of species designation. They concluded that the study of Fennessey et al. (2016) should only be regarded as one perspective on giraffe taxonomy. On the lack of power of the nuclear dataset used, Fennessey et al. (2017) argued that "Compared to microsatellite data, DNA sequences allow estimating divergence times". Fennessey et al. (2016), however, did not estimate population divergence times, only sequence divergence times, which, incidentally, can also be estimated with microsatellites (e.g. Hey 2010). The response by Bercovitch et al. (2017) also highlighted different criteria for species delimitation than Fennessey et al. (2016, 2017). Whilst Fennessey et al. (2016, 2017) advocate diagnosability using (neutral) genetic markers as the primary criteria for species delineations, Bercovitch et al. (2017) placed a greater emphasis on phenotypic and behavioural characteristics. They stated that: "Coat color patterns are linked to specific gene complexes with mutations leading to variation subject to natural selection… Phenotypic traits regulate mating patterns and sexual selection that establish a foundation for the recognition species concept".

Ultimately, Fennessey et al. (2016) used limited genetic data to detect genetic structure and sequence divergence criteria, which were then equated with species divergence by applying the PSC. However, the process of lineage sorting under plausible demographic and selection models was not considered, nor their influence in the context of the limited number of markers used. A follow-up study using a larger set of nuclear markers has since been carried out, which confirms that gene-flow between the four proposed species

is very low (Winter et al. 2018). However, it appears that in this situation the argument is predominantly of an ontological nature, and so may not have run its course yet.

Orang-utan

Nater et al. (2017) recently described the genomic diversity of a population of orangutans from the species' southernmost range limit in Sumatra (Batang Toru). They concluded that the Batang Toru population was sufficiently distinct to warrant being named a new species. This conclusion was based on morphometric, behavioural and genomic evidence from 33 to 37 individuals (the morphological analysis could only use a single Batang Toru specimen). Using Approximate Bayesian Computation modelling of demography, it was estimated that the northern Sumatra population split from the older Batang Toru ~ 3.4 million years ago (mya), but maintained gene flow until 10–20 thousand year ago (kya). The authors also point out that there are many instances of ongoing gene flow between taxa that are recognised as distinct, well-established species. In light of this, Nater et al. (2017) use the species definition that describes species as "a population (or group of populations) with fixed heritable differences from other such populations (or groups of populations)", effectively invoking the PSC.

The morphological evidence which led to the conclusion of a new orang-utan species was based on a single specimen from the population in question (and genomic evidence based on two). Any criticisms of the validity/robustness of this conclusion could be centred around the question of whether a single specimen can be considered representative of the whole population. Nater et al. (2017) point out that numerous species have been identified based on a single type specimen in the past. Based on genomics, the authors were able to show that these two orangutan populations had fixed heritable differences with an estimated termination of gene-flow from/to the proposed new species 10–20 kya. Yet, Nater et al. (2017) did not assess if these SNPs were associated with adaptive differences between the populations. Thus, although Nater et al. (2017) used genomics to enhance their power to apply the PSC with greater resolution, they did not use it to attempt to understand the speciation process in any mechanistic sense. The conclusions reached by Nater et al. (2017) has not been accepted by all in the scientific community, particularly by proponents of the BSC (e.g. https://whyevolutionistrue.wordpress.com/2017/11/03/a-new-species-of-orangutan-i-doubt-it/). Nater et al. (2017) pointed out that determining if these populations are reproductively isolated or not is not possible, due to their allopatric distribution. One potential solution that was not used by Nater et al. (2017) is the Tobias criteria (Tobias et al. 2010). This uses sympatric species pairs to set thresholds for delineating allopatric taxa. It seems likely that despite the large number of features investigated, and analytical methods applied, this approach will still fall short of the expectations of many proponents of the BSC.

In short, the orang-utan paper represents a case in which a large panel of the genomic tools available have been applied to address the question of population divergence. While presumably adding detailed information about the historical processes, it does not attempt to analyse adaptive differences, nor to answer whether maintaining these two populations of orang-utan as separate would maximize the adaptive potential going forward.

Finless porpoise

Zhou et al. (2018) investigated speciation in finless porpoises, which have traditionally been classified as a single species, *Neophocaena phocaenoides*. Finless porpoises exist as three geographic populations or subspecies, two marine (Indo-Pacific) and one freshwater population (Yangtze River). Zhou et al. (2018) identified several candidate genes related to hypoxia that show strong evidence of directional selection. They also estimated divergence of the Yangtze River population at 5000–40,000 years ago. These findings led them to conclude that "significant population differentiation, lack of gene flow, and unique adaptive divergence in the Yangtze finless porpoise make it clear that the Yangtze finless porpoise is genetically and reproductively isolated from its marine counterpart and thus represents an incipient species".

The main aspect that differentiates the porpoise case study from that of the orangutan is the term "unique adaptive divergence". By identifying selection signatures in several candidate genes that are the result of diversifying selection to two different ecosystems, Zhou et al. (2018) found plausible mechanistic evidence for an instance of incipient speciation. Whilst the orang-utan study by Nater et al. (2017) showed phenotypic differences between the two proposed species, no evidence was presented to demonstrate that this divergence was adaptive, and therefore driving speciation. This highlights the issue that, although genomic methods for identifying selection in natural populations has advanced considerably over recent years, it is still challenging to do this with limited numbers of samples.

Darwin's finches

Lamichhaney et al. (2017) documented a remarkable example of hybrid speciation from its origin to reproductive isolation in a hybrid between two Darwin's finch species (*Geospiza fortis* and *G. conirostris*). This hybrid lineage was shown to breed endogamously from the second generation onwards, with transgressive segregation of bill morphology, a trait that is known to be under strong selective pressure in these species.

This study demonstrates that reproductive isolation can occur rapidly, in as little as three generations. This species classification was therefore based on reproductive isolation of the new hybrid finch lineage from its parent lineages, aka the BSC.

Prima facie, the question of a new species of Darwin's finch seems very simple: These species exist in sympatry, and were observed to stop interbreeding, a situation clearly fulfilling the criteria of distinct species under the BSC. However, Hill and Zink (2018) firstly notes that three to four generations may not be enough time to determine if the new lineage is ephemeral or not, and secondly that phenotypic differences observed may be highly plastic. The conclusions of Lamichhaney et al. (2017) are strengthened by the fact that they also investigated the genetic basis for bill dimension, a morphological trait that is implicated in driving ecological success and reproductive isolation of the big bird lineage. By observing correlations between the ALX1 and HMGA2 loci with morphometrics, they were able to use genomics to provide evidence for genetic adaptation to a new environment. It seems unlikely that the level of observational evidence that they used will be practical for most wild species, a common criticism of the practicality of the BSC (Amato and Russello 2014). However, there are genomic approaches that can bypass these challenges for many taxa. For example, relatives, pedigrees, and local ancestry tracts can be identified so that reproductive isolation over the last few generations can be inferred from genetic data (e.g. as carried out in humans, Ko and Nielsen 2017). This could serve as an alternative to observational studies.

This is not to say that there are not conceptual criticisms that can be made of the BSC regardless of how it is operationalized [e.g. related to instances of viable hybrids between organisms well-recognised to be different species (Nater et al. 2017)]. As discussed earlier, a full discussion of this is beyond the scope of this manuscript, however, genomic tools are at least allowing us to be able to better quantify and understand the relevance of these instances (even when we only have low coverage data or few individuals, Abbott et al. 2016).

Genomic and other data increasingly show that these hybridization and introgression events can no longer be classed as a rare or insignificant: they are now being recognised as both common and important evolutionary mechanisms, including sometimes being implicated in the adaptive advantages to a newly colonised environment (e.g. invertebrates, Pogson 2016; plants; Ru et al. 2016; and vertebrates; Barbato et al. 2017).

The role of hybridisation in species designation

Hybridization is ubiquitous in nature. Sixteen percent of bird species (Ottenburghs et al. 2015), 6% of European mammals and at least 25% of vascular plants (Mallet 2005) are thought to hybridise. Ravinet et al. (2018) investigated signatures of divergence and introgression in a species pair: The Pacific Ocean three-spined stickleback (*Gasterosteus aculeatus*) and the Japan Sea stickleback (*G. nipponicus*). These are well-recognised as different species that have sympatric distributions and crosses showing male hybrid sterility (Kitano et al. 2007). However, despite the high differentiation, relatively large divergence time (0.68–1 mya) and hybrid sterility, ongoing gene-flow and localised introgression could nonetheless be detected (maintained in small regions within the genome). Although the authors are not making a new species claim, this observation of introgression despite the considerable divergence time is highly relevant to the speciation question.

This situation provides challenges for both the PSC and the BSC. How infrequent do hybridization events have to occur before we consider two biological entities to be different species? Does it make a difference if such hybridization is sex-biased? How does regional variation in hybridization rates influence this inference? The BSC currently has no answer to these questions. Likewise, for the PSC, "fixed heritable differences" will be immediately mixed in hybrid individuals. Therefore, temporal or spatial variation in hybridization could lead to transient or spatially varying species classifications.

Due to the increasing recognition of the pervasiveness of hybridization and introgression among recognised species, they are becoming important phenomena to consider when making taxonomic decisions. The idea that hybridization may play an important role in evolution was initially explored by botanists and appears to be particularly important for plants, with approximately 10% of plant species thought to hybridize (Yakimowski and Rieseberg 2014). Hybridization is also particularly common in invasive species (Ellstrand and Schierenbeck 2000), likely due to hybridization allowing adaptive introgression of beneficial traits between the taxa (Martin et al. 2005, 2006). However, widespread hybridization is not limited to plants and has played an important role in the adaptive radiation of *e.g.* Heliconius butterflies (Dasmahapatra et al. 2012). These butterflies are of particular interest in speciation research because of their huge diversity, with varying rates of hybridization (Van Belleghem et al. 2017). Their genomes contain what has become known as "genomic islands of divergence" (Nadeau et al. 2012). Originally identified in *Anopheles* mosquitos (Turner et al. 2005), the origin and role of these islands was originally interpreted as regions of selection and reduced introgression between divergent populations, although it is increasingly being realised that there are processes other than population divergence that may lead to these patterns (Cruickshank and Hahn 2014; Wolf and Ellegren 2016).

Hybridization complicates taxonomy when we consider that speciation rates, and levels of subsequent hybridization

vary considerably between taxa. The proposed new species of Darwin's finch described above showed transgressive segregation in bill morphology and was ecologically successful. This ongoing finch radiation is predominantly based on a behavioural trait (i.e. mate choice). Finches imprint on features of their parents early in life, and choose mates based on bill size and shape, and body size and song. The driving force behind the speciation events here is therefore a complex mating behaviour. While these adaptive traits (at least in the case of bill dimensions) are correlated with detectable genetic variation, it is their effect on the behaviour phenotype that is relevant for reproductive isolation and species designation in these taxa. It seems fair to assume that if the observational data were available, this situation would be representative for most taxa with complex mating behaviour. However, this is in stark contrast to many other taxonomic groups, which can take far longer to develop reproductive isolation. For example, hybridization in marine invertebrates may be extreme. One study found hybridization between two cryptic species of sea squirt (*Ciona intestinalis*) with an average synonymous sequence divergence of 14.4% (Roux et al. 2013). Rates of introgression in *Ciona* were relatively low, variable among loci, and unidirectional, consistent with a situation of multiple genetic incompatibilities throughout the genome, suggesting that genetic incompatibility was developing, albeit very slowly. It would be interesting to use genomics to investigate signatures of selection in these *Ciona* populations, to see the extent to which adaptation can be detected, and how it reflects the taxonomy.

Previously, we might have written off these examples of extreme hybridization as being exceptional, however this explanation is becoming more difficult to abide. As we can see from the stickleback example above (Ravinet et al. 2018), the phenomenon is not limited to invertebrates. In fact, whole genome data are detecting instances of introgression in many species and in unprecedented detail. For example, most non-African humans have 1–2% Neanderthal ancestry (Green et al. 2010; Prüfer et al. 2014), and a number of human populations have Denisovan ancestry that is thought to have adaptive significance for adaptation to extreme altitude (Reich et al. 2010; Meyer et al. 2012; Prüfer et al. 2014). Such patterns of introgression are mirrored in non-human primates, with evidence of multiple occurrences between bonobos and chimpanzees during the past 550,000 years (De Manuel et al. 2016).

These observations complicate the matter of species delineation, because they suggest that complete reproductive isolation can be withheld for extremely long periods of time in some taxa (in the case of *Ciona*, for greater than three million years of divergence in isolation). It could be argued that this is just the BSC impartially reflecting the variable speciation rates that occur in nature, however some taxonomists (e.g. with well-known mammalian groups) clearly find such observations problematic as these instances do not tend to be reflected taxonomically (e.g. between brown and polar bears, coyotes and wolves). Some concepts may regard hybridisation as a "consequence", while others think of it as a defining characteristic. However, hybridisation does not only complicate species designation for the latter. Hybrids may not initially seem relevant to the PSC, but hybrid zones between two different taxa diagnosed using the PSC would create a gradient of alleles, such that the sampling scheme (across the geographic space as well as the genome) and population comparison chosen would dictate whether taxa would be diagnosed as different. This present a challenge, not only for diagnosing different units, but also for describing what those things are from an ontological point of view.

Are the species concepts operational in the genomic era?

There are therefore challenges in operationalizing species concepts, but is this more the case for some rather than others? And how has genomic data facilitated operationalisation for each concept? The PSC is easier to test in most cases, and Groves (2013) argued that "the PSC offers the only criterion for species recognition that is testable, as a scientific proposition should be." However, it might be questioned in what sense the PSC is testable. And if so, is it the only species delineation approach that is?

As argued by Groves (2013), the PSC is "testable", however when we do this we must be cautious that we are not engaging in an oversimplification. The application of a testable threshold does not represent progress if that threshold does not reflect the label that we are trying to establish. Genetic differentiation among populations can be greatly influenced by demography, including changes in population size. Genetic structure has been observed to considerably decrease among brown bear populations (*Ursus arctos*) in just 1.5 generations (Hagen et al. 2015), and genetic structure substantially increased over only 11 years (approx. five generations) in Coachella Valley fringe-toed lizards (*Uma inornata*; Vandergast et al. 2016). These examples are not intended to demonstrate that speciation does not occur over short time periods, but simply that genetic divergence and population structure may be highly transient, which many people would argue should not be the case for speciation. Genomics allows for a huge increase in the power to detect population structure because of the much larger number of loci available. This has the effect of enabling the identification of very fine-scale population genetic structure, and consequently more 'fixed heritable differences' between populations. 'Splitters' would presumably interpret this added genomic information as an increase in power of detecting incipient speciation, whereas 'lumpers' would presumably

interpret these as 'type 1 error' species. In this regard then whether genomics has revolutionised our ability to identify new species depends on the species concept being applied. Genomics has also allowed for a huge increase in the power to describe demographic histories (e.g. Nater et al. 2017), and this information is important to present alongside that of genetic structure when making a species claim, so that that claim can be assessed in its full context.

All but one of the examples discussed here have used genetic structure as a part of their evidence, however it is notable that the porpoise (Zhou et al. 2018), orangutan (Nater et al. 2017) and stickleback (Ravinet et al. 2018) studies also include demographic analysis, whereas the giraffe study (Fennessy et al. 2016) did not. Genetic structure does not distinguish between isolation and migration and so is very difficult to interpret on its own. In addition, the first three studies above used a methodology and dataset that enabled them to estimate genetic structure that was representative of the whole genome. As we can see from the stickleback example (Ravinet et al. 2018), and the discussion on introgression above, using genetic structure based on a small number of loci can be misleading: even species with high genomic divergence may have introgressed regions that will give a very different perspective of the taxonomy—and even without gene flow incomplete lineage sorting can generate a high proportion of "wrong" gene trees (Jarvis et al. 2014).

An understanding of population structure can be important for conservation, but it is important to understand its limitations. Frankham et al. (2012) argued that species delineations need to be relevant to the point at which populations have/have not become reproductively isolated (which is not necessarily related to genetic structure), in order for them to minimise the risk of inbreeding and outbreeding depression and maximise the benefits of gene-flow. These arguments led the authors to recommend that only substantial reproductive isolation be used to define species (for outbreeding sexual organisms) in conservation. Amato and Russello (2014) commented on this paper, with their main critique being the difficulty of operationalising the BSC. Frankham (2014) countered that reproductive isolation generally arises from adaptation to different environments and/or outbreeding depression caused by fixed chromosomal differences, both of which can be detected (albeit requiring a more technically challenging approach than a structure analysis). They stated that "Divergence should be protected when it reflects adaptive differences, but countered when it threatens populations." The authors were therefore arguing that the BSC is be a better proxy for adaptive potential than the PSC. It is important to note that this argument is predominantly based on the BSC being a better tool for recognising conservation units, and therefore is not addressing its ontological relevance. Nonetheless, adaptive potential *is* important if we want to conserve populations that are able to adapt to changes in their environment. However, is it true that the BSC preserves adaptive potential better, and if so, are there limits and/or exceptions to this?

Which species concept best conserves adaptive potential?

Adaptation to novel ecological opportunities is one of the main drivers of speciation (Van Belleghem et al. 2017), and predicting the capacity of taxonomic groupings to respond to changing environments is therefore crucial to their conservation (Eizaguirre and Baltazar-Soares 2014). The Darwin's finch example above is a clear demonstration of the potential of hybridization to produce a population with unique adaptive potential. However, this hybridization and introgression may have a confounding influence on species delineations (particularly for the BSC), which is exacerbated when we also consider the adaptive advantage that introgressed genes may bring. This process, adaptive introgression, poses a challenge to the claim that the BSC is a good proxy for adaptive potential. Even very low levels of introgression can have a large effect on the adaptive potential of the recipient population; adaptive genetic variation has the potential to move to high frequencies very quickly in a population (Maynard Smith and Haigh 2008). In addition, the adaptive potential of the introgressed material may vary between the donor and recipient populations, depending on factors such as population size and selection regime. Therefore, in some situations, taxa designated by the BSC (even when allowing for very low levels of introgression) may be reflective of adaptive differences between them (e.g. the adaptive differences in the Darwin's finch example). However, in many situations it will not. For example, it seems highly likely that the two distinct populations of sea squirts (*Ciona*) (Roux et al. 2013) have accumulated considerable adaptive differentiation in their three million years of divergence in isolation, regardless of the fact that gene-flow has now been re-established. This gene-flow would preclude these as separate species under the BSC, and therefore (unlike with the finches) the taxonomy would not reflect the adaptive differences between populations/species. Hence, the BSC will better represent adaptive differentiation in some comparisons than in others, and this may be biased towards taxonomic groups with particular life-history traits. It should also be noted that this is no less the case for the PSC. If our goal is to conserve adaptive potential in an unbiased way across all taxa then this is a crucial point to consider. Many scientists argue that maximizing phylogenetic diversity will indirectly capture functional diversity (Vane-Wright et al. 1991; Faith 1992; Winter et al. 2013). However, a recent study by Mazel et al. 2018 has shown that phylogenetic diversity does not reliably capture functional diversity.

This raises the question of why not simply measure adaptive potential directly? Genomics is starting to allow us to do this. For example, Zhou et al. (2018) identified evidence of selective sweeps in a number of genomic regions across the porpoise genome using a method that looks for distinctive patterns of allele frequencies along a chromosome (Nielsen et al. 2005). Other commonly used methods for detecting selection include: (1) Identification of extended haplotypes that are at, or near fixation in a subset of individuals (Sabeti et al. 2007), (2) Outlier methods that compare a model based on including versus excluding selection (Foll and Gaggiotti 2008), (3) Attempts to identify correlations between SNPs and environmental variables (Coop et al. 2010). In the porpoise example, Zhou et al. (2018) found regions that have a plausible link to morphological characteristics that differentiate the two proposed incipient species. Applying these methods has the benefit of not requiring the assumption that adaptive differences are related to reproductive isolation or genetic structure, which, as described above, may be inaccurate. It should be noted however that tracking adaptive changes using genomics is challenging for many traits, especially those that have low heritability or are highly polygenic (Hoffmann et al. 2017). However, it is often hard to convincingly demonstrate selection on a given region of the genome as in many cases it is only the regions undergoing strong haplotypic selection that will be detected in the analyses discussed above. Furthermore, demonstrating past selection may not necessarily be associated with contemporary or future adaptive potential of a genome/genomic region, given that selection pressures are dynamic. Finally, even if a genomic region can be identified as being under selection, determining the specific "cause" of this pressure can be highly challenging, particularly for non-model organisms.

We have argued that some species concepts may be more applicable (in terms of relating to adaptive potential) to some taxa than others. For example, reproductive isolation may be a useful criterion in the case of Darwin's finches, since it aligns with the behavioural, morphological and ecological differences between populations. For organisms like sea squirts, genetic distance and differentiation may be a better reflection of the differences that have accumulated over long periods of temporal and spatial isolation. The relationship between adaptive potential and species concept therefore seems to depend on the taxa being investigated. This does not necessarily mean that these are not good criteria, independently, for defining species. However, it certainly complicates conservation strategies that aim to maximise evolutionary potential, especially when only one is considered at a time. We would therefore caution against focusing on a single species concept, especially when the taxa in question are of conservation concern. In this situation it is important to be very clear about which concepts are being invoked, and how the evidence presented supports them. It is important to incorporate multiple lines of evidence into taxonomic decisions (which is increasingly being done; Schlick-Steiner et al. 2010) however, this evidence can now theoretically be provided by entirely by genomics: (1) morphological evidence can be identified via differentiation in developmental and structural genes, (2) biogeographic evidence can be provided using sophisticated genome-scale modelling, (3) behavioural differences can be inferred by identifying genes associated with behaviour, mate-choice, and also by detecting sex-biased demography, (4) ecological evidence is available in the form of genomic signatures of selection to environmental factors, (5) reproductive compatibility can be observed as sex chromosome compatibility/incompatibility, chromosomal structure, and epigenomic transmission. In lieu of a definitive conclusion as to the most appropriate species concept to be used, best practice would be to investigate as many of the above lines of evidence as possible, and to apportion ones confidence in a species designation based on the combined weight of all of them. Recently, Kitchener et al. (2017) introduced the concept of a 'traffic light' system for evaluating the strength of evidence of the above five categories of species differentiation, which may provide a pragmatic approach to evaluating genomic data in specific definition if applied sensibly.

One thing that both 'splitters' and 'lumpers' seem to agree on is that it is preferable that conservation decisions are based on sound scientific evidence. Any 'planning blight' due to taxonomic uncertainty can be detrimental to conservation, and renders decisive action more difficult. However, while we still have some way to go before genomic techniques reach their full potential as a diagnostic tool for species delineation, if the ultimate goal of conservation is to preserve adaptive potential, genomics is now allowing us to gain a better understanding of this in wild populations. A pragmatic approach could be to use genomic tools to characterise adaptive potential regardless of the species concept, or even without invoking a species concept at all. However, answering the question of whether and to what extent such studies *should* focus on adaptive potential is a separate challenge.

Acknowledgements MWB gratefully acknowledges the influence of Professor HC Macgregor, who died 22/7/2018, on his contribution to this article.

References

Abbott RJ, Barton NH, Good JM, Schaefer NK, Shapiro B, Green RE (2016) Detecting hybridization using ancient DNA. Mol Ecol 25(11):2398–2412

Amato G, Russello MA (2014) Operationalism matters in conservation: comments on Frankham et al. (2012). Biol Conserv 170:332–333. https://doi.org/10.1016/j.biocon.2009.02.031

Avise JC, Ball RM (1990) Principles of genealogical concordance in species concepts and biological taxonomy. In: Futuyama D, Antonovics J (eds) Oxford surveys in evolutionary biology. Oxford University Press, Oxford, pp. 45–67

Baker RJ, Bradley RD (2006) Speciation in mammals and the genetic species concept. J Mammal 87(4):643–662. https://doi.org/10.1644/06-MAMM-F-038R2.1

Barbato M, Hailer F, Orozco-Terwengel P, Kijas J, Mereu P, Cabras P, Mazza R, Pirastru M, Bruford MW (2017) Genomic signatures of adaptive introgression from European mouflon into domestic sheep. Sci Rep 7(1):1–13. https://doi.org/10.1038/s41598-017-07382-7

Baum DA, Shaw KL (1995) Genealogical perspectives on the species problem. In: Hoch PC, Stevenson AG (eds) Experimental and molecular approaches to plant biosystematics. Monographs in systematics. Missouri Botanical Garden, St. Louis, pp. 289–303

Bercovitch FB, Berry PSM, Dagg A, Deacon F, Doherty JB, Lee DE, Mineur F, Muller Z, Ogden R, Seymour R, Shorrocks B, Tutchings A (2017) How many species of giraffe are there? Curr Biol Elsevier 27(4):R136–R137. https://doi.org/10.1016/j.cub.2016.12.039

Brown DM, Brenneman RA, Koepfli K, Pollinger JP, Milá B, Georgiadis NJ, Jr EEL, Grether GF, Jacobs DK, Wayne RK, Louis EE, Grether GF, Jacobs DK (2007) Extensive population genetic structure in the giraffe. 13, pp. 1–13. https://doi.org/10.1186/1741-7007-5-57

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 185(4):1411–1423. https://doi.org/10.1534/genetics.110.114819

Cracraft J (1997) Species concepts in systematics and conservation biology – an ornithological viewpoint. In: Claridge MF, Dawah HA, Wilson MR (eds) Species: The Units of Biodiversity. Chapman & Hall, London, pp 325–339

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol 23(13):3133–3157. https://doi.org/10.1111/mec.12796

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin AV, Hughes DST, Ferguson LC, Martin SH, Salazar C, Lewis JJ, Adler S, Ahn SJ, Baker DA, Baxter SW, Chamberlain NL, Ritika C, Counterman BA, Dalmay T, Gilbert LE, Gordon K, Heckel DG, Hines HM, Hoff KJ, Holland PWH, Jacquin-Joly E, Jiggins FM, Jones RT, Kapan DD, Kersey P, Lamas G, Lawson D, Mapleson D, Maroja LS, Martin A, Moxon S, Palmer WJ, Papa R, Papanicolaou A, Pauchet Y, Ray DA, Rosser N, Salzberg SL, Supple MA, Surridge A, Tenger-Trolander A, Vogel H, Wilkinson PA, Wilson D, Yorke JA, Yuan F, Balmuth AL, Eland C, Gharbi K, Thomson M, Gibbs RA, Han Y, Jayaseelan JC, Kovar C, Mathew T, Muzny DM, Ongeri F, Pu LL, Qu J, Thornton RL, Worley KC, Wu YQ, Linares M, Blaxter ML, ffrench-Constant RH, Joron M, Kronforst MR, Mullen SP, Reed RD, Scherer SE, Richards S, Mallet J, Mc Millan WO, Jiggins CD (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487(7405):94–98. https://doi.org/10.1038/nature11041

De Queiroz K (1998) The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations. In: Howard DJ, Berlocher SH (eds) Endless forms. Species and speciation. Oxford University Press, Oxford, pp 57–75

de Queiroz K (1999) The general lineage concept of species and the defining properties of the species category. In: Wilson RA (ed) Species: new interdisciplinary essays. MIT Press, Cambridge, pp. 49–89

De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, Schmidt JM, Heredia-Genestar JM, Benazzo A, Barbujani G, Peter BM, Kuderna LFK, Casals F, Angedakin S, Arandjelovic M, Boesch C, Kühl H, Vigilant L, Langergraber K, Novembre J, Gut M, Gut I, Navarro A, Carlsen F, Andrés AM, Siegismund HR, Scally A, Excoffier L, Tyler-Smith C, Castellano S, Xue Y, Hvilsom C, Marques-Bonet T (2016) Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354(6311):477–481. https://doi.org/10.1126/science.aag2602

Eizaguirre C, Baltazar-Soares M (2014) Evolutionary conservation-evaluating the adaptive potential of species. Evol Appl 7(9):963–967. https://doi.org/10.1111/eva.12227

Eldredge N, Cracraft J (1980) Phylogenetic patterns and the evolutionary process. Columbia University Press, New York

Ellstrand NC, Schierenbeck KA (2000) Hybridization as a stimulus for the evolution of invasiveness in plants? Proc Natl Acad Sci USA 97(13):7043–7050bib>

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61:483 1–10

Fennessy J, Bidon T, Reuss F, Vamberger M, Fritz U, Janke A, Fennessy J, Bidon T, Reuss F, Kumar V, Elkan P, Nilsson MA, Vamberger M (2016) Multi-locus analyses reveal four giraffe species instead of one report multi-locus analyses reveal four giraffe species instead of one. Curr Biol 26:1–7

Fennessy J, Winter S, Reuss F, Kumar V, Nilsson MA, Vamberger M, Fritz U, Janke A (2017) Response to "How many species of giraffe are there?". Curr Biol Elsevier 27(4):R137–R138. https://doi.org/10.1016/j.cub.2016.12.039

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics 180(2):977–993. https://doi.org/10.1534/genetics.108.092221

Frankham R (2014) Species concepts for conservation—reply to Russello and Amato. Biol Cons 170:334–335. https://doi.org/10.1016/j.biocon.2009.01.015

Frankham R, Ballou JD, Dudash MR, Eldridge MDB, Fenster CB, Lacy RC, Mendelson JR, Porton IJ, Ralls K, Ryder OA (2012) Implications of different species concepts for conserving biodiversity. Biol Conserv 153:25–31

Gippoliti S, Cotterill FPD, Zinner D, Groves CP (2017) Impacts of taxonomic inertia for the conservation of African ungulate diversity: an overview. Biol Rev (April). https://doi.org/10.1111/brv.12335

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano H, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S (2010) A draft sequence of the Neandertal genome. Science (New York), 328(5979):710–722. https://doi.org/10.1126/science.1188021

Groves C, Grubb P (2011) Ungulate taxonomy. The Press, Johns Hopkins University, Baltimore
Groves CP (2013) The nature of species: a rejoinder to Zachos et al. Mamm Biol 78(1):7–9. https://doi.org/10.1016/j.mambio.2012.09.009
Grummer JA, Bryson RW, Reeder TW (2014) Species delimitation using Bayes factors: simulations and application to the Sceloporus scalaris species group (Squamata: Phrynosomatidae). Syst Biol 63:119–133
Hagen SB, Kopatz a, Aspi J, Kojola I, Eiken HG (2015) Evidence of rapid change in genetic structure and diversity during range expansion in a recovering large terrestrial carnivore. Proc R Soc B 282(1807):20150092–20150092. https://doi.org/10.1098/rspb.2015.0092
Heller R, Frandsen P, Lorenzen ED, Siegismund HR (2013) Are there really twice as many bovid species as we thought? Syst Biol 62(3):490–493. https://doi.org/10.1093/sysbio/syt004
Hey J (2010) Isolation with migration models for more than two populations. Mol Biol Evol 27(4):905–920. https://doi.org/10.1093/molbev/msp296
Heywood (2010) Explaining patterns in modern ruminant diversity: contingency or constraint? Biol J Lin Soc 99:657–672
Hill GE, Zink RM (2018) Hybrid speciation in birds, with special reference to Darwin's finches. J Avian Biol 49(9):.e01879
Hoffmann AA, Sgrò CM, Kristensen TN (2017) Revisiting adaptive potential, population size, and conservation. Trends Ecol Evol 32(7):506–517. https://doi.org/10.1016/j.tree.2017.03.012
Isaac NJB, Mallet J, Mace GM (2004) Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol Evol 19(9):464–469. https://doi.org/10.1016/j.tree.2004.06.004
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-Núñez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jønsson KA, Johnson W, Koepfli K-P, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alström P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MTP, Zhang G (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215):1320 LP–1331
Kingdon J (2013) The Kingdon field guide to African mammals. A&C Black, London
Kitano J, Mori S, Peichel CL (2007) Phenotypic divergence and reproductive isolation between sympatric forms of Japanese threespine sticklebacks. Biol J Lin Soc 91(4):671–685. https://doi.org/10.1111/j.1095-8312.2007.00824.x
Kitchener AC, Breitenmoser-Würsten Ch, Eizirik E, Gentry A, Werdelin L, Wilting A, Yamaguchi N, Abramov AV, Christiansen P, Driscoll C, Duckworth JW, Johnson W, Luo S-J, Meijaard E, O'Donoghue P, Sanderson J, Seymour K, Bruford M, Groves C, Hoffmann M, Nowell K, Timmons Z, Tobe S (2017) A revised taxonomy of the Felidae. The final report of the Cat Classification Task Force of the IUCN/SSC Cat Specialist Group. Cat News Spec Issue 11:80 pp
Ko A, Nielsen R (2017) Composite likelihood method for inferring local pedigrees. PLOS Genet 13(8):e1006963
Kutschera VE, Bidon T, Hailer F, Rodi JL, Fain SR, Janke A (2014) Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol Biol Evol 31(8):2004–2017. https://doi.org/10.1093/molbev/msu186
Lamichhaney S, Han F, Webster MT, Andersson L, Grant BR, Grant PR (2017) Rapid hybrid speciation in Darwin's finches. Science 7120(September):eaao4593
Mailund T, Munch K, Schierup MH (2014) Lineage sorting in apes. Annu Rev Genet 48(1):519–535. https://doi.org/10.1146/annurev-genet-120213-092532
Malaney JL, Cook JA (2013) Using biogeographical history to inform conservation: the case of Preble's meadow jumping mouse. Mol Ecol 22(24):6000–6017. https://doi.org/10.1111/mec.12476
Mallet J (2005) Hybridization as an invasion of the genome. Trends Ecol Evol 20(5):229–237
Martin NH, Bouck AC, Arnold ML (2005) Loci affecting long-term hybrid survivorship in *Louisiana irises*: implications for reproductive isolation and introgression. Evol Int J Org Evol 59(10):2116–2124. https://doi.org/10.1111/j.0014-3820.2005.tb00922.x
Martin NH, Bouck AC, Arnold ML (2006) Detecting adaptive trait introgression between *Iris fulva* and *I. brevicaulis* in highly selective field conditions. Genetics 172(4):2481–2489. https://doi.org/10.1534/genetics.105.053538
Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge MF et al. (eds.), Species: the units of biodiversity, Chapman & Hall, London, pp. 381–424
Maynard Smith J, Haigh J (2008) The hitch-hiking effect of a favourable gene. Genet Res 89(5–6):391–403. https://doi.org/10.1017/S0016672308009579
Mayr E (1942) Animal species and evolution. Columbia University Press, Cambridge
Mayr E (1963) Systematics and the origin of species. Columbia University Press, New York
Mazel F, Pennell MW, Cadotte MW, Diaz S, Riva GVD, Grenyer R, Leprieur F, Moores AO, Mouillot D, Tucker CM, Pearse WD (2018) Prioritizing phylogenetic diversity captures functional diversity unreliably. Nat Commun 9:2888
McCarthy et al (2012) Financial costs of meeting global biodiversity conservation targets: current spending and unmet needs. Science 338(6109):946–949. https://doi.org/10.1126/science.1229803
Meyer M, Kircher M, Gansauge M, Li H, Racimo F, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Slatkin M, Reich D (2012) A high-coverage genome sequence from an archaic denisovan individual a high-coverage genome sequence from an archaic denisovan individual a high-coverage genome sequence from an archaic denisovan individual. Science (New York) 222(2012):1–14. https://doi.org/10.1126/science.1224344
Morrison WR, Lohr JL, Duchen P, Wilches R, Trujillo D, Mair M, Renner SS (2009) The impact of taxonomic change on conservation: does it kill, can it save, or is it just irrelevant? Biol Cons 142(12):3201–3206. https://doi.org/10.1016/j.biocon.2009.07.019
Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, Ffrench-Constant RH, Blaxter ML, Mallet J, Jiggins CD (2012) Genomic islands of divergence in hybridizing heliconius butterflies identified by large-scale targeted sequencing. Philos Trans R Soc B 367(1587):343–353. https://doi.org/10.1098/rstb.2011.0198
Nater A, Mattle-greminger MP, Nurcahyo A, Nowak MG, Meijaard E (2017) Morphometric, behavioral, and genomic evidence. Curr Biol 27:1–12. https://doi.org/10.1016/j.cub.2017.09.047
Nielsen R, Williamson S, Kim Y, Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for

selective sweeps using SNP data Genomic scans for selective sweeps using SNP data. Genome Res. https://doi.org/10.1101/gr.4252305

Orozco-terWengel P, Corander J, Schötterer C (2011) Genealogical lineage sorting leads to significant, but incorrect Bayesian multilocus inference of population structure. Mol Ecol 20:1108–1121

Ottenburghs J, Ydenberg RC, Van Hooft P, Van Wieren SE, Prins HHT (2015) The Avian Hybrids Project: gathering the scientific literature on avian hybridization. Ibis 157(4):892–894

Ottenburghs J, Megens H-J, Kraus RHS, van Hooft P, van Wieren SE, Crooijmans RPMA, Ydenberg RC, Groenen MAM, Prins HHT (2017) A history of hybrids? Genomic patterns of introgression in the True Geese. BMC Evol Biol 17:201

Pogson GH (2016) Studying the genetic basis of speciation in high gene flow marine invertebrates. Curr Zool 62(6):643–653. https://doi.org/10.1093/cz/zow093

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, De Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505(7481):43–49. https://doi.org/10.1038/nature12886

Ravinet M, Yoshida K, Shigenobu S, Toyoda A, Fujiyama A, Kitano J (2018) The genomic landscape at a late stage of stickleback speciation: high genomic divergence interspersed by small localized regions of introgression. PLOS Genet. https://doi.org/10.1371/journal.pgen.1007358

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Pääbo S (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468(7327):1053–1060. https://doi.org/10.1038/nature09710

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *ciona intestinalis* species. Mol Biol Evol 30(7):1574–1587. https://doi.org/10.1093/molbev/mst066

Ru D, Mao K, Zhang L, Wang X, Lu Z, Sun Y (2016) Genomic evidence for polyphyletic origins and interlineage gene flow within complex taxa: a case study of *Picea brachytyla* in the Qinghai-Tibet Plateau. Mol Ecol 25(11):2373–2386. https://doi.org/10.1111/mec.13656

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds D, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura PKH, Kawaguchi Y, Kitamoto T, Morizono T, Nagashima T, Ohnishi A, Sekine Y, Tanaka A, Tsunoda T, Deloukas T, Bird P, Delgado CP, Dermitzakis M, Gwilliam ET, Hunt R, Morrison S, Powell J, Stranger D, Whittaker BE, Bentley P, Daly DR, de Bakker MJ, Barrett PIW, Chretien J, Maller YR, McCarroll J, Patterson S, Pe'er N, Price I, Purcell A, Richter S, Sabeti DJ, Saxena P, Sham R, Stein PC, Krishnan LD, Smith L, Tello-Ruiz AV, Thorisson MK, Chakravarti GA, Chen A, Cutler PE, Kashuk DJ, Lin CS, Abecasis S, Guan GR, Li W, Munro HM, Qin ZS, McVean DJ, Auton G, Bottolo A, Cardin L, Eyheramendy N, Freeman S, Marchini C, Myers J, Spencer S, Stephens C, Donnelly M, Cardon P, Clarke LR, Evans G, Morris DM, Weir AP, Johnson BS, Mullikin T, Sherry JC, Feolo ST, Skol M, Zhang A, Matsuda H, Fukushima I, Macer Y, Suda DR, Rotimi E, Adebamowo CN, Ajayi C, Aniagwu I, Marshall T, Nkwodimmah P, Royal C, Leppert CDM, Dixon MF, Peiffer M, Qiu A, Kent R, Kato A, Niikawa K, Adewole N, Knoppers IF, Foster BM, Clayton MW, Watkin EW, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913–918

Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH (2010) Integrative taxonomy: a multisource approach to exploring biodiversity. Annu Rev Entomol 55:421–438. https://doi.org/10.1146/annurev-ento-112408-085432

Simpson GG (1951) The species concept. Evolution 5(4):285–298

Simpson GG (1961) Principles of animal taxonomy. Columbia University Press, New York

Sonsthagen SA, Wilson RE, Chesser RT, Pons JM, Crochet PA, Driskell A, Dove C (2016) Recurrent hybridization and recent origin obscure phylogenetic relationships within the 'white-headed' gull (*Larus* sp.) complex. Mol Phylogenet Evol 103:41–54. https://doi.org/10.1016/j.ympev.2016.06.008

Sukumaran J, Knowles LL (2017) Multispecies coalescent delimits structure, not species. PNAS 114(7):1607–1612. https://doi.org/10.1073/pnas.1607921114

Tobias JA, Seddon N, Spottiswoode CN, Pilgrim JD, Fishpool LD, Collar NJ (2010) Quantitative criteria for species delimitation. Ibis 152(4):724–746

Toussaint et al (2016) Bayesian Poisson tree processes and multispecies coalescent models shed new light on the diversification of Nawab butterflies in the Solomon Islands (Nymphalidae, Charaxinae, Polyura). Zool J Linn Soc 178:241–256

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol 3(9):1572–1578. https://doi.org/10.1371/journal.pbio.0030285

Van Belleghem SM, Rastas P, Papanicolaou A, Martin SH, Arias CF, Supple MA, Hanly JJ, Mallet J, Lewis JJ, Hines HM, Ruiz M, Salazar C, Linares M, Moreira GRP, Jiggins CD, Counterman BA, McMillan WO, Papa R (2017) Complex modular architecture around a simple toolkit of wing pattern genes. Nat Ecol Evol 1(3):1–12. https://doi.org/10.1038/s41559-016-0052

Vandergast AG, Wood DA, Thompson AR, Fisher M, Barrows CW, Grant TJ (2016) Drifting to oblivion? Rapid genetic differentiation in an endangered lizard following habitat fragmentation and drought. Divers Distrib 22(3):344–357. https://doi.org/10.1111/ddi.12398

Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect?—systematics and the agony of choice. Biol Conserv 55:235–254

Wheeler QD, Platnick NI (2000) The phylogenetic species concept (sensu Wheeler & Platnick). In: Wheeler QD, Meier R (eds) Species concepts and phylogenetic theory: a debate. Columbia University Press, New York, pp 55–69

Wiley EO (1978) The evolutionary species concept. Syst Zool 27(1):17–26. https://doi.org/10.2307/2412809

Winter M, Devictor V, Schweiger O (2013) Phylogenetic diversity and nature conservation: where are we? Trends Ecol Evol 28(4):199–204

Winter S, Fennessy J, Janke A (2018) Limited introgression supports division of giraffe into four species. Ecol Evol. https://doi.org/10.1002/ece3.4490

Wolf JBW, Ellegren H (2016) Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. https://doi.org/10.1038/nrg.2016.133

Yakimowski SB, Rieseberg LH (2014) The role of homoploid hybridization in evolution: a century of studies synthesizing genetics and ecology. Am J Bot 101(8):1247–1258. https://doi.org/10.3732/ajb.1400201

Yang Z (2015) The BPP program for species tree estimation and species delimitation. Curr Zool 61:5

Zachos FE (2013) Taxonomy: species splitting puts conservation at risk. Nature 494(7435):35. https://doi.org/10.1038/494035c

Zachos FE, Apollonio M, Bärmann EV, Festa-Bianchet M, Göhlich U, Habel JC, Haring E, Kruckenhauser L, Lovari S, McDevitt AD, Pertoldi C, Rössner GE, Sánchez-Villagra MR, Scandura M, Suchentrunk F (2013) Species inflation and taxonomic artefacts-A critical comment on recent trends in mammalian classification. Mamma Biol 78(1):1–6. https://doi.org/10.1016/j.mambio.2012.07.083

Zhou X, Guang X, Sun D, Xu S, Li M, Seim I, Jie W, Yang L, Zhu Q, Xu J, Gao Q, Kaya A, Dou Q, Chen B, Ren W, Li S, Zhou K, Gladyshev VN, Nielsen R, Fang X, Yang G (2018) Population genomics of finless porpoises reveal an incipient cetacean species adapted to freshwater. Nat Commun 9(1):1–8. https://doi.org/10.1038/s41467-018-03722-x

nature COMMUNICATIONS

Corrected: Author correction

ARTICLE

DOI: 10.1038/s41467-018-07070-8 OPEN

Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits

Arthur Gilly[1], Daniel Suveges[1], Karoline Kuchenbaecker[1,2,3], Martin Pollard[1,4], Lorraine Southam[1,5], Konstantinos Hatzikotoulas[1,6], Aliki-Eleni Farmaki[7,8], Thea Bjornland[9], Ryan Waples[10], Emil V.R. Appel[11], Elisabetta Casalone[12], Giorgio Melloni[13], Britt Kilian[1], Nigel W. Rayner[1,5,14], Ioanna Ntalla[15], Kousik Kundu[1,16], Klaudia Walter[1], John Danesh[1,17,18], Adam Butterworth[17,18,19], Inês Barroso[1], Emmanouil Tsafantakis[20], George Dedoussis[8], Ida Moltke[10] & Eleftheria Zeggini[1,6]

The role of rare variants in complex traits remains uncharted. Here, we conduct deep whole genome sequencing of 1457 individuals from an isolated population, and test for rare variant burdens across six cardiometabolic traits. We identify a role for rare regulatory variation, which has hitherto been missed. We find evidence of rare variant burdens that are independent of established common variant signals (*ADIPOQ* and adiponectin, $P = 4.2 \times 10^{-8}$; *APOC3* and triglyceride levels, $P = 1.5 \times 10^{-26}$), and identify replicating evidence for a burden associated with triglyceride levels in *FAM189B* ($P = 2.2 \times 10^{-8}$), indicating a role for this gene in lipid metabolism.

[1] Department of Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom. [2] Division of Psychiatry, University College of London, London W1T 7NF, United Kingdom. [3] UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom. [4] Department of Medicine, Addenbrooke's Hospital, University of Cambridge, Hills Road, Cambridge CB2 0QQ, United Kingdom. [5] Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom. [6] Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg D-85764, Germany. [7] Department of Health Sciences, College of Life Sciences, University of Leicester, Leicester LE1 6TP, United Kingdom. [8] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens 176-71, Greece. [9] Department of Mathematical Sciences, Norwegian Institute of Science and Technology, Trondheim 7491, Norway. [10] The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark. [11] Section for Metabolic Genetics, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen 2200, Denmark. [12] Human Genetics Foundation, University of Torino, Torino IT-10126, Italy. [13] Department of Biomedical Informatics, Harvard Medical School, Boston 02115 MA, USA. [14] Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Old Road, Headington, Oxford OX3 7LE, United Kingdom. [15] William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, United Kingdom. [16] Department of Haematology, Cambridge Biomedical Campus, University of Cambridge, Long Road, Cambridge CB2 0PT, United Kingdom. [17] The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, University of Cambridge, Cambridge CB1 8RN, United Kingdom. [18] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Wort's Causeway, University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, United Kingdom. [19] British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom. [20] Anogia Medical Centre, Anogia 740 51, Greece. These authors contributed equally: Arthur Gilly, Daniel Suveges, Karoline Kuchenbaecker. Correspondence and requests for materials should be addressed to E.Z. (email: Eleftheria@sanger.ac.uk)

Genome-wide association studies have gleaned substantial insights into the genetic architecture of complex traits. The contribution of common-frequency variants to complex traits has been well-documented, and progress in understanding the role of low frequency variation has also gained considerable traction. However, the role of rare variants in the genetic architecture of medically-relevant complex traits remains less well-understood, and the allelic architecture of complex trait association signals has not yet been fully resolved. Rare variant association studies have so far mainly focussed on exonic regions[1], and in whole-genome sequencing studies the optimal analytical approach for rare regulatory variants remains an open question[2]. Population-scale deep whole genome sequencing can capture genetic variation across the entire allele frequency spectrum traversing the coding and non-coding genome. In addition, population isolates offer increased power gains in detecting associations involving rare and low-frequency variants[3].

Here, to improve our understanding of the role of rare variants, we perform cohort-wide deep whole genome sequencing of 1457 individuals from a deeply-phenotyped, isolated population from Crete, Greece (the HELIC-MANOLIS cohort[4–6]) at an average depth of 22.5× (Supplementary Fig. 1), capturing 98% of true single nucleotide variants (SNVs) (Methods and Supplementary Fig. 2). The population genetics characteristics of HELIC-MANOLIS have been studied, and indicate an effective population size of $N_e = 6242$ and an approximate time of divergence of 1100 years from the general Greek population[4,7]. We address open questions on whole genome sequencing study design, analysis and interpretation, and identify burdens of coding and regulatory rare variants associated with cardiometabolic traits.

Results

Effect of sequencing depth. Comparing whole genome sequencing at various depths ranging from 15× to 30× (Methods), we find that 96.4% of singletons, 97.9% of doubletons and 97.6% of variants called using 30× sequencing are recapitulated at 22.5x depth. Genotype accuracy (as measured by r^2) is 99.7% for 22.5× depth and 98.5% for 15× depth, suggesting that increases between 15× and 30× translate into marginal improvements in both call rate and quality of very rare SNVs (Fig. 1, Supplementary Fig. 3 and Methods). We find that false discovery rates and genotype accuracy are substantially more dependent on sequencing depth for INDELs than for SNVs (Fig. 1).

Landscape of sequence variation. Following quality control (QC), we call 24,163,896 non-monomorphic SNVs and INDELs, 97.9% of which are biallelic. 14,281,180 (60.31%) of the biallelic SNVs are rare (minor allele frequency [MAF] < 0.01); 3,103,273 (13.1%) are low-frequency (MAF 0.01–0.05); and 6,292,726 (26.57%) are common (MAF > 0.05). We call 8,294 non-monomorphic variants annotated as loss-of-function (LoF) with low-confidence (LC)[8], and 438 variants annotated as LoF with high-confidence (HC) (Supplementary Fig. 4). On average, each individual carries 405 (standard deviation $\sigma = 19$) LC LoF variants and 31 ($\sigma = 6$) HC LoF variants, compared to 149 LoF variants per sample in a whole genome sequencing study of 2636 Icelanders[9]. 0.6 and 1% of HC and LC LoF carrier genotypes are homozygous, respectively. INDELs are significantly more frequent among LoF variants, with 53.2 and 76% in the low-confidence and high-confidence sets, respectively, compared to 13.5% genome-wide. We observe an enrichment of rare variants among the coding and splice variant categories in MANOLIS (one-sided exact binomial $P = 9.5 \times 10^{-16}$), and we recapitulate this in an independent dataset of 3724 individuals with whole genome sequencing from the UK-based INTERVAL cohort[10](Fig. 2). We also observe a lower rate of singletons compared to the general Greek population and the INTERVAL cohort ($P \approx 10^{-167}$ and $P < 10^{-200}$, respectively, one-sided empirical P-value) (Methods and Supplementary Fig. 5), in keeping with the isolated nature of this Cretan population. Among the 5,102,175 novel biallelic variants (not present in gnomAD[11] or Ensembl release 84[12]), 4,394,678 are SNVs, and the majority are rare (Supplementary Fig. 6).

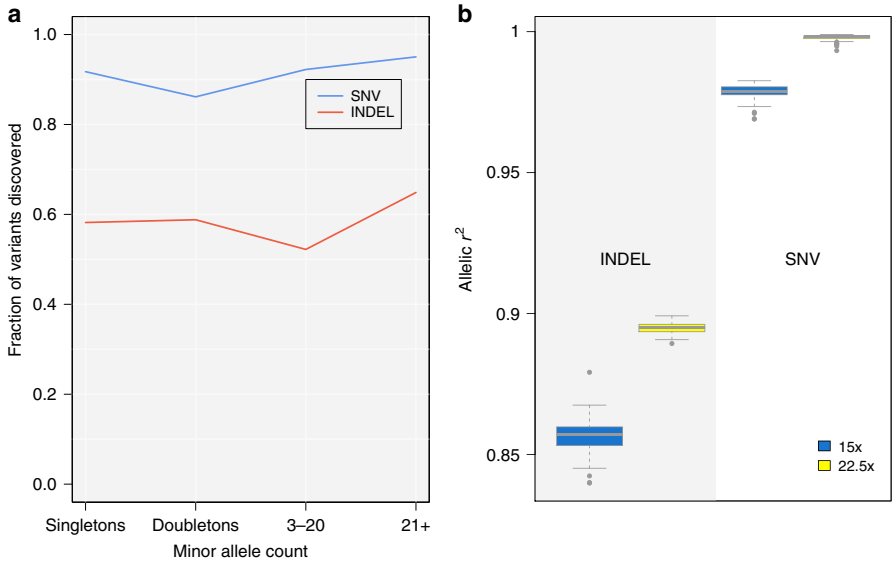


Fig. 1 Variant discovery and quality in WGS data from 100 samples. **a** variant discovery rate in 22.5×; **b** allelic r^2 for SNVs and INDELs in both 15× and 22.5× calls. Depth is downsampled randomly from 30×. INDEL: insertion/deletion. SNV: single nucleotide variant. Boxes represent the interquartile range. Bold horizontal lines in boxplots represent the median, the whiskers extend to 1.5 times the interquartile range, and grey dots represent outliers outside the whisker range

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-07070-8 ARTICLE

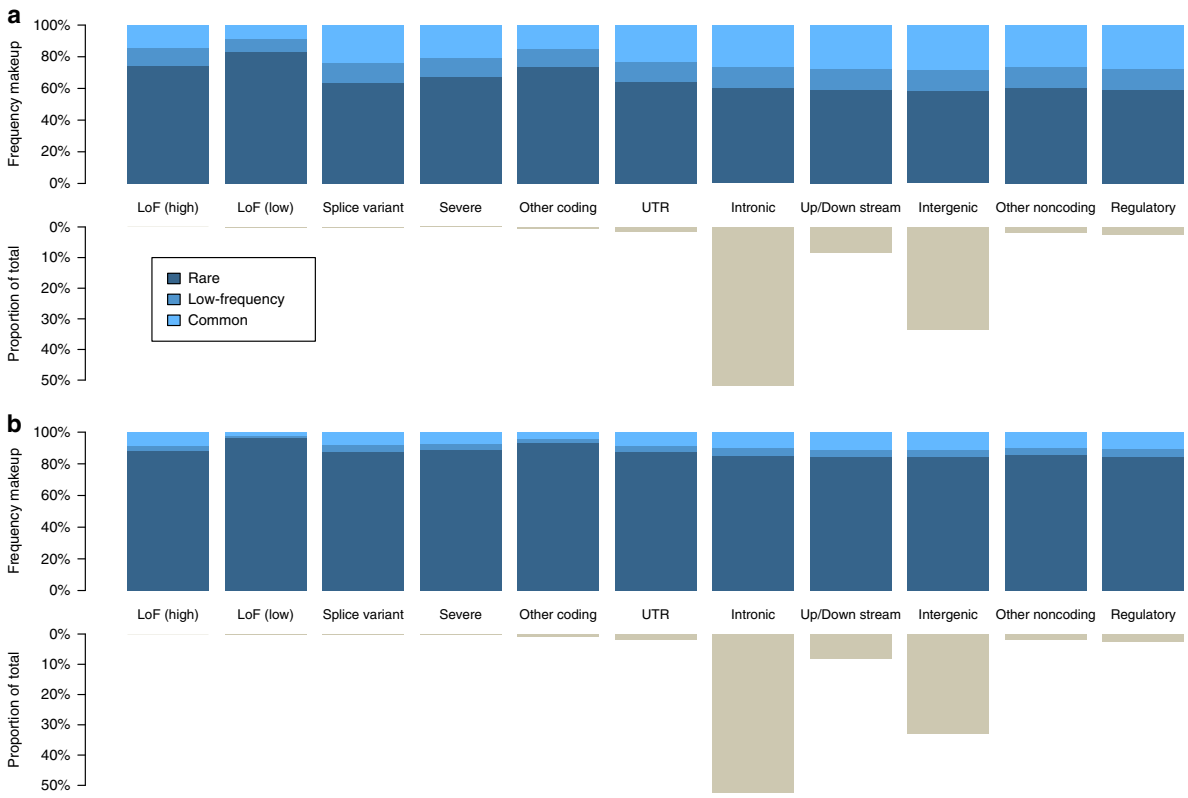


Fig. 2 Variant count proportions and minor allele frequency bin by functional class. **a**, **b** Data is shown for MANOLIS (**a**) and INTERVAL (**b**). Functional classes are derived from the Ensembl VEP consequences as detailed in Supplementary Table 6. The number of intergenic variants is likely to be an underestimate due to Ensembl's most severe consequence annotation. For each panel, the bottom half represents the proportion of variants in each class relative to the total number of variants, the upper half represents the frequency makeup of variants in each class

Refinement of parameters for rare variant burden testing. We carried out genome-wide rare variant burden analyses for six medically-relevant traits: serum adiponectin, bilirubin, gamma-glutamyltransferase, low- and high-density lipoprotein, and triglyceride levels. As choice of genomic region, variant selection and weighting remain open questions for rare variant analysis, we benchmark 10 approaches using different regions of interest (exonic, exonic and regulatory, and regulatory only), variant inclusion and weighting methods (Methods; Supplementary Table 1). Overall, association statistics correlate highly within three distinct clusters (Supplementary Fig. 7). Among exonic-only analyses, rare variant tests that only include unweighted high-consequence variants cluster separately from those in which variants are weighted according to their functionality scores. The third cluster encompasses all tests that include regulatory variants. Neither the variant weighting scheme nor the transformation used for adjusting the weights has a notable influence on the results.

Rare variant burden discovery. In total, twenty burden signals exceed the study-wide significance threshold of 2.0×10^{-7} (Supplementary Fig. 8), arising from four independent genes. Providing proof-of-principle, we identify association of a burden of loss-of-function variants with blood triglyceride and high-density lipoprotein levels in the *APOC3* gene (Fig. 3.a, Supplementary Data File 1)[5,13]. The strongest signal arises when the splice-donor variant rs138326449 (minor allele count (MAC) = 38, minor allele frequency (MAF) = 0.013) and the stop-gained variant rs76353203 (MAC = 62, MAF = 0.022) are included in the analysis ($P = 1.6 \times 10^{-26}$). We replicate the association of a burden of rare coding *APOC3* variants with triglyceride levels in INTERVAL, in which we identify a burden of 25 exonic variants ($P = 3.1 \times 10^{-6}$) (Supplementary Data File 2). This is driven by rs138326449 and rs187628630, a rare 3' UTR variant (MAF = 0.008), with a two-variant burden $P = 9.0 \times 10^{-7}$. rs138326449 is the only loss-of-function variant in *APOC3* present in this cohort, and is four times rarer than in MANOLIS ($MAF_{INTERVAL} = 0.003$ vs $MAF_{MANOLIS} = 0.013$).

We detect a new association of triglyceride levels with rare variants in the *FAM189B* gene (Fig. 3.b, Supplementary Data File 1). The burden association ($P = 1.5 \times 10^{-7}$) is driven by two independent novel splice variants: chr1:155251911 G/A (human genome build 38, MAC = 3, $P = 8.2 \times 10^{-6}$) and chr1:155254079 C/G (MAC = 2, $P = 6.04 \times 10^{-4}$). In both cases, the minor allele is associated with increased triglyceride levels (effect size $\beta = 2.59$ units of standard deviation, $\sigma = 0.57$ and $\beta = 2.40$ $\sigma = 0.69$, respectively). Both variants exhibit high quality scores (VQSLOD > 19), high sequencing read depth (24× and 26.5×, respectively) and no missingness. A further novel splice region variant (chr1:155251496 T/C) and a stop gained variant (rs145265828), both singletons, were also included in the analysis; however their contribution to the burden is insignificant (burden $P = 2.2 \times 10^{-8}$ when excluding them). We replicate evidence for a burden signal at *FAM189B* in the INTERVAL cohort ($P = 9.3 \times 10^{-3}$) (Supplementary Data File 2), which includes two stop gained variants with one driving the association: chr1:155250417 (rs749626426, MAC = 2, $\beta = 1.96$ $\sigma = 0.70$, $P = 5.4 \times 10^{-3}$). In

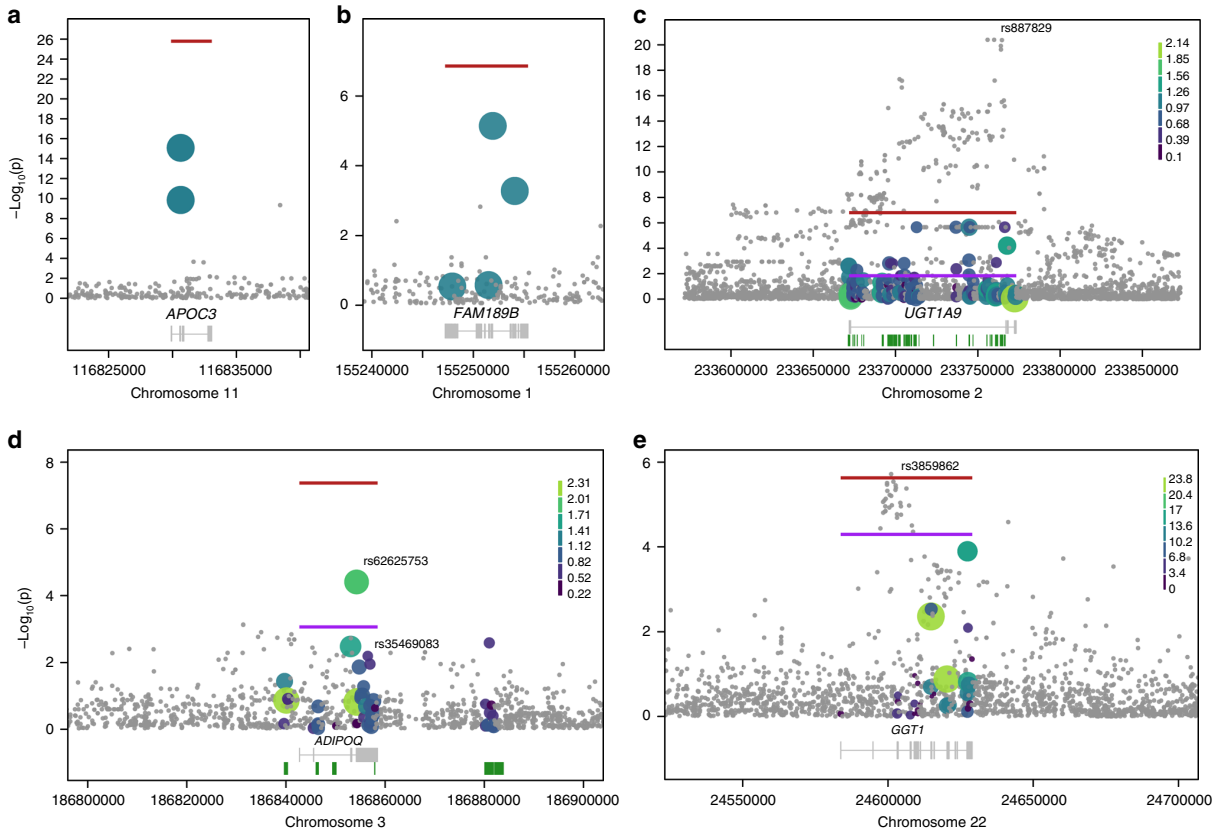


Fig. 3 Regional association plots for burdens in *APOC3, FAM189B, UGT1A9, ADIPOQ,* and *GGT1*. Red lines denote the burden *P*-value and extend over the tested gene. Purple lines indicate the conditioned *P*-value for the variant described in the text (variants rs887829, rs62625753, and rs3859862 in **c**, **d**, **e**. respectively). Small grey dots indicate single-point *P*-value for variants in the region not included in the test. Larger coloured dots represent variants included in the test, with size and colour proportional to the score used in the most significant test (Supplementary Table 1 and Supplementary Data File 1). When no weights are applied (**a** and **b**), included variants are coloured blue-green. In the gene track below the regional plots, green bars below the gene, if present, denote regulatory regions associated with the gene which were used to include variants in the burden. These are present only for genes where regulatory regions were included in the burden

keeping with the discovery dataset, the disruptive minor allele is associated with increased triglyceride levels. The two novel splice-region variants discovered in MANOLIS are not present in either the INTERVAL study or in a compendium of 123,136 exomes and 15,496 whole genomes assembled as part of the gnomAD project[11]. *FAM189B* has not been previously associated with blood lipid levels.

We find evidence of a low frequency and rare variant burden association with bilirubin levels in the *UGT1A9* gene (Fig. 3.c, Supplementary Data File 1). This association arises from the analyses including exonic and regulatory variants ($P = 1.9 \times 10^{-8}$), and from the analyses including regulatory variants only ($P = 7.2 \times 10^{-8}$). We find evidence for association in the exon plus regulatory region burden analysis in the INTERVAL replication cohort ($P = 1.7 \times 10^{-45}$, Supplementary Data File 2). A common variant in the first intron of *UGT1A9* (rs887829, MAF = 0.28, $\beta = 0.426$ $\sigma = 0.04$, two-sided score test $P = 4.0 \times 10^{-21}$ in the MANOLIS cohort) has previously been associated with bilirubin levels[14,15]. As expected, genotype correlation between rs887829 and each of the low-frequency and rare variants included in the burden is low ($r_{max}^2 = 0.1$). The rs887829 signal is not attenuated when conditioning on carrier status for the two main drivers of the burden (single-point score test $P_{conditional} = 4.5 \times 10^{-21}$), or when conditioning on the number of rare alleles carried per individual ($P_{conditional} = 4.0 \times 10^{-21}$). The evidence for association with the rare variant burden in *UGT1A9* is substantially reduced when conditioned on rs887829 (burden $P_{conditional} = 0.0146$). Conversely, the two-variant signal for the two main burden drivers is attenuated from $P = 1.4 \times 10^{-7}$ to $P_{conditional} = 7.0 \times 10^{-3}$ when conditioning on rs887829, indicating that it likely recapitulates part of a signal driven by a known common-variant association in the region.

We identify an association of adiponectin levels with low-frequency and rare variants in the *ADIPOQ* gene (Fig. 3.d, Supplementary Data File 1). The evidence for association is stronger for exonic and regulatory variants combined ($P = 4.2 \times 10^{-8}$) than in either the regulatory-only ($P = 0.19$) or exon-only ($P = 2.0 \times 10^{-6}$) analyses, suggesting a genuine contribution of both classes of variants to the burden. The missense variant rs62625753 (MAF = 0.031, two-sided score test $P = 4.0 \times 10^{-5}$) contributes to the burden signal and is predicted to be damaging. The strength of association for the burden is reduced, but not entirely attenuated, when conditioned on rs62625753 ($P_{conditional} = 8.9 \times 10^{-4}$), indicating that it is not singly driven by this variant. rs35469083 (MAF = 0.044) also contributes to the burden, and is an expression quantitative trait locus (eQTL) for *ADIPOQ* in visceral adipose tissue (minor allele associated with decreased gene expression). rs62625753 and rs35469083 have

consistent directions of effect, with the minor alleles associated with reduced adiponectin levels, in keeping with their functional consequences on the gene (two-variant burden $P = 4.8 \times 10^{-7}$). No common-variant signal for adiponectin levels is present in this region in our dataset. The burden signal remains significant upon conditioning on the genotypes of all variants with previous associations for adiponectin, type 2 diabetes or obesity that are polymorphic in MANOLIS (Supplementary Table 2).

In addition to the four genes that meet study-wide significance, we find gamma-glutamyltransferase levels to be suggestively associated with a burden of low frequency and rare exonic variants in the gamma-glutamyltransferase 1 (*GGT1*) gene ($P = 2.3 \times 10^{-6}$) (Fig. 3.e, Supplementary Data File 1). A previously-reported, common-variant association is also present in an intron of this gene (rs3859862, MAF = 0.46, two-sided score test $P = 1.9 \times 10^{-6}$). The burden signal in *GGT1* is maintained when conditioning on rs3859862 ($P_{conditional} = 5.1 \times 10^{-5}$), suggesting that rare variants be independently contributing to this established association. Similarly, the single-point association at rs3859862 conditioned on carrier status for all rare variants included in the burden is not attenuated ($P_{conditional} = 2.8 \times 10^{-5}$), a result recapitulated by conditioning the same variant on the number of rare alleles carried per individual ($P_{conditional} = 1.8 \times 10^{-5}$), providing evidence for an independent rare variant signal at this locus.

Signatures of selection. We surveyed the genomic loci with evidence of rare variant burden signals for signatures of recent or ongoing positive selection in the MANOLIS cohort, using integrated haplotype scores (iHS)[16]. Previous studies have shown that an elevated fraction of SNVs with $|iHS| > 2$ in a genomic region is a signature of recent or ongoing selection and notably, we find that 32% of the SNVs in *FAM189B* have an iHS score above 2, placing it in the top 5% of all genes analysed (96.7th percentile). This result is robust across several definitions of the genomic region representing the genes (95.6th-98.3th percentile) and to conditioning on gene length (94.6th percentile) (Supplementary Table 3). To further investigate this potential signature of selection in *FAM189B*, we examined the extent to which the allele frequencies in *FAM189B* differ between the MANOLIS cohort and the 1000 Genomes CEU population sample using weighted mean F_{ST}. Like with iHS, *FAM189B* lies in the top 5% of all genes analysed across several definitions of the genomic regions (Supplementary Table 4).

Discussion

In this work, we have whole genome sequenced 1457 individuals from the HELIC-MANOLIS cohort at an average depth of 22.5×. We describe the genomic variation landscape in this special population, discover 5.1 million novel variants, and perform rare variant burden testing across the entire genome for medically-relevant biochemical traits.

We empirically address several open whole genome sequencing study design and analysis questions. Through a downsampling approach, we demonstrate that it is possible to achieve near-perfect sensitivity and quality for rare SNV calling and genotyping with half the depth, and at substantially lower cost, compared to 30× sequencing. This observation does not extend to INDELs, for which depth increases above 15× can result in a 15% increase in genotype quality and a 40% increase in true positive rate.

Defining the genomic regions in which to select variants, filtering strategies and variant weighting schemes constitute unresolved challenges in whole genome sequence-based studies. We find that association signal profiles of tests including regulatory region variants differ markedly from other scenarios, with some signals being driven by this variant class. Further, signal strength differs substantially between analyses that include high-severity consequence exonic variants only, and those in which all exonic variants are weighted according to their predicted consequence. We find that, as a rule, variant and functional unit selection, rather than weighting scheme, plays the largest role in association testing.

We identify a role for rare regulatory variants in the allelic architecture of complex traits. It is therefore important to leverage the whole genome sequence nature of the study data, and not to restrict analyses to coding variation only. We observe congruent directions of effect among regulatory and coding rare variants in burden signals that combine both classes of variation, for example across eQTL and damaging missense variants in the *ADIPOQ* gene that are together associated with adiponectin levels.

We discover replicating evidence for association of a rare variant burden with triglyceride levels at a locus not previously linked with the trait. *FAM189B* (Family With Sequence Similarity 189 Member B), also known as *COTE1* or *C1orf2*, codes for a membrane protein that is widely expressed, including in adult liver tissue[17]. Expression of *FAM189B* has been found to be correlated with endogenous SREBP-1 activation in vitro[18]. Sterol-regulatory element binding proteins (SREBPs) control the expression of genes involved in fatty acid and cholesterol biosynthesis, therefore indicating a mechanism by which *FAM189B* could be involved in lipid metabolism. We found *FAM189B* to contain an elevated fraction of SNVs with $|iHS|>2$, a potential signature of recent positive selection. Furthermore, *FAM189B* is in the top 5% of all genes in terms of population differentiation (F_{ST}) between the MANOLIS cohort and the 1000 Genomes Project CEU sample, which is consistent with selection having happened in MANOLIS. This is particularly interesting in the context of this population, which has a high animal fat content diet[6], and for which loss of function variants in *APOC3* have risen in frequency compared to the general population and confer a cardioprotective effect[19,20]. For the same reason, it is interesting to note that *FAM189B* has not previously been reported to be under selection in other populations[21]. However, we caution that although *FAM189B* is in the top 5% of all genes for both iHS and F_{ST}, it is not an extreme outlier for either, suggesting that it could be a false positive or that the selection has not acted strongly enough or for long enough to leave more than subtle signatures in the haplotype structure and allele frequencies in the gene. It is also possible that selection has acted on several rare alleles making the signature more complex than simple directional selection.

We replicate the *FAM189B* association in an independent dataset with deep whole genome sequence data, in which the disruptive rare alleles are also associated with the same trait in the same direction. Across the board, we replicate all burden signals for which replication cohort trait measurements are available. We find that allelic heterogeneity is prevalent, partly due to the rare nature of the variants contributing to the burdens, and partly due to the distinct population genetics characteristics of the discovery and replication sets. Perhaps as a consequence, the outcome of variant filtering and weighting was quite sensitive to the study population, and all the burdens reported here replicated strongest in slightly different testing conditions from the discovery, although in the same broad functional class. The association in *FAM189B* was discovered when including exonic variants with a relaxed severity threshold, whereas it replicated in the LoF-only analysis. Similarly, the *APOC3* signal was discovered in the LoF-only analysis but replicated in the CADD-weighted exonic analysis. These findings have important consequences for defining replication in sequence-based studies of rare variants, and

highlight the importance of defining replication at the locus level rather than the variant level for burden signals.

We demonstrate pervasive allelic heterogeneity at complex trait loci, and identify exonic and regulatory rare variant associations at established signals. We find multiple instances of burden signals that remain independent of colocalising common variant signals, and one instance of burden signal attenuation when conditioning on the established common variant association. Within the power constraints of the study, we do not find evidence for synthetic association at established signals, i.e., there is no evidence for multiple rare variants at a locus accounting for a common variant association.

The discovery of rare variant burden associations with a modest sample size has been made possible due to the special population genetics characteristics of the isolated cohort under study. Rare variant signals, such as the ones discovered in *APOC3* and *FAM189B* in MANOLIS, are driven by variants with severe consequences that are rarer or absent in cosmopolitan populations. This demonstrates that the well-rehearsed power gains conferred by isolated cohorts in genome-wide association studies[3] extend to whole genome sequence-based rare variant association designs.

Our findings indicate that deep whole genome sequencing at scale will be required to enable exhaustive description of the rare variant burden landscape in a population. For example, in the case of the *FAM189B* signal, low-depth sequencing (1× depth) of 1239 MANOLIS samples[22] misses one of the two burden-driving variants (chr1:155251911, MAC = 3). Similarly, genome-wide genotyping coupled to dense imputation of the same samples does not capture the variants driving the burden signal identified here through deep whole genome sequencing[23].

Our findings provide evidence for a role of low-frequency and rare, regulatory and coding variants in complex traits, and highlight the complex nature of locus-specific architecture at established and newly emerging signals. We anticipate that larger-scale, cohort-wide, deep whole genome sequencing initiatives will substantially further contribute to our understanding of the genetic underpinning of complex traits.

Methods

Ethics and informed consent statement. In the TEENAGE study, prior to recruitment all study participants gave their verbal assent along with their parents'/guardians' written consent forms. The study was approved by the Institutional Review Board of Harokopio University and the Greek Ministry of Education, Lifelong Learning and Religious Affairs. The MANOLIS study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant. The INTERVAL study was approved by the Cambridge South Research Ethics Committee and informed consent was obtained from every participant.

Sequencing. For MANOLIS, genomic DNA (500 ng) from 1482 samples was sheared to a median insert size of 500 bp and subjected to standard Illumina paired-end DNA library construction. Adapter-ligated libraries were amplified by 6 cycles of PCR and subjected to DNA sequencing using the HiSeqX platform (Illumina) according to manufacturer's instructions. For TEENAGE, one hundred samples from the general Greek population were sequenced, as well as the Genome in a Bottle NA12878 sample. Sample identity checks were performed using Fluidigm and aliquots prepared. These aliquots underwent library preparation using the standard HiSeqX method. Size selection was performed to target 350 base pairs. Sequencing was performed on the Sanger Institute's Illumina HiSeqX plat-form with a target depth of 30x and PhiX spike-in.

Evaluation of sequencing accuracy at various depths. Reads from the NA12878 were downsampled to several read depths (from 5× to 30×) using the -s option of samtools view, aligned and processed through GATK Variant Quality Score Recalibrator. They were then compared to Genome in a Bottle (GIAB) 0.2 calls to extract the true positive rate (Supplementary Fig. 2). At 22.5×, true positive rates are 98% for SNVs and 76% for INDELs.

Comparison with the general Greek population. We compared variant callsets in MANOLIS to a dataset of 100 samples from the Greek general population (TEENAGE study), for which an identical sequencing protocol was used. The average depth in the TEENAGE study was 32.1×. We downsampled the individual BAMs to 22.5× and 15× using the -s option of samtools based on the average depth of the TEENAGE dataset, then performed variant calling using GATK HaplotypeCaller v3.3 (https://github.com/mp15/af_analysis) and filtering using GATK Variant Quality Score Recalibrator. The downsampled and original datasets were then compared using bcftools stats to extract allelic r-squared (Fig. 1.b.). For the 22.5× dataset, we compared variant overlap with bcftools isec (Fig. 1.a. and Supplementary Fig. 3).

Rare variant counts in MANOLIS, TEENAGE and INTERVAL. Since sample sizes differ between the three datasets, we randomly subsampled the larger dataset to a matching size for each pairwise comparison. We used these resampled datasets to build empirical distributions for rare variant counts in the larger dataset, and compared it to counts in the smaller dataset. TEENAGE ($n = 100$) was smaller compared to MANOLIS ($n = 1482$), so we drew 1000 sets of 100 samples from the MANOLIS study for the comparison. We counted 270,916 singletons and 61,690 doubletons in TEENAGE, compared with a median of 179,100 (one-sided $P = 1.4 \times 10^{-94}$ from a fitted normal distribution) and 75,280 (one-sided $P = 3.0 \times 10^{-19}$ from a fitted normal distribution), respectively, in MANOLIS (Supplementary Fig. 5a,b.). For n = 100, singletons correspond to MAF < 0.005 and doubletons to 0.005 < MAF < 0.1.

For the INTERVAL ($n = 3742$) comparison, MANOLIS was the smaller dataset, so we resampled 500 sets of 1482 samples from the INTERVAL cohort and counted variants up to MAC = 29 (MAF = 0.01). The increased resolution provided by this larger sample size shows that rare variant counts are greater in the cosmopolitan population below MAC = 4 (MAF = 0.0013), but greater in the isolate for 0.0013 < MAF < 0.1, consistent with our coarser observation in TEENAGE (Supplementary Fig. 5.c).

For p-values of singleton counts, empirical quantiles cannot be computed for such large deviations from the mean. We fitted a normal distribution to singleton counts, and computed the theoretical quantile corresponding to the observed count in the smaller cohort.

Variant calling. Basecall files for each lane were transformed into unmapped BAMs using Illumina2BAM, marking adaptor contamination and decoding barcodes for removal into BAM tags. PhiX control reads were mapped using BWA Backtrack and were used to remove spatial artefacts. Reads were converted to FASTQ and aligned using BWA MEM 0.7.8 to the 1000 Genomes hs37d5 (for NA12878) and hg38 (GRCh38) with decoys (HS38DH) (for TEENAGE) references. The alignment was then merged into the master sample BAM file using Illumina2BAM MergeAlign. PCR and optical duplicates are marked using biobambam markduplicates and the files were archived in CRAM format.

Per-lane CRAMs were retrieved and reads pooled on a per-sample basis across all lanes to produce library CRAMs; these were each divided in 200 chunks for parallelism. GVCFs were generated using HaplotypeCaller v.3.5 from the Genome Analysis Toolkit (GATK) for each chunk. All chunks were then merged at sample level, samples were then further combined in batches of 150 samples using GATK CombineGVCFs v.3.5. Variant calling was then performed on each batch using GATK GenotypeGVCFs v.3.5. The resulting variant callsets were then merged across all batches into a cohort-wide VCF file using bcftools concat.

Quality control. Variant-level QC was performed using the Variant Quality Score Recalibration tool (VQSR) from the Genome Analysis Toolkit (GATK) v. 3.5-0-g36282e4[24], using a tranche threshold of 99.4% for SNPs, which provided an estimate false positive rate of 6%, and a true positive rate of 95%. For INDELs, we used the recommended threshold of 1%. For sample-level QC, we made extensive use of a previously described[23] GWAS dataset in 1175 overlapping samples. Four individuals failed sex checks, 8 samples had low concordance ($\hat{\pi}$<0.8) with chip data, 11 samples were duplicates, and 12 samples displayed traces of contamination (Freemix score from the verifyBamID suite[25]>5%). In case of sample duplicates, the sample with highest quality metrics (depth, freemix and chipmix score) was kept. As contamination and sex mismatches were correlated, a total of 25 individuals were excluded ($n = 1457$). No further samples were excluded based on depth, heterozygosity, transition/transversion (Ti/Tv) rate, missingness or ethnicity. No rare or low-frequency variant (MAF < 5%) was excluded based on the Hardy-Weinberg equilibrium test at $P = 1.0 \times 10^{-5}$. We filtered out 14% of variants with call rates < 99%.

Genetic relatedness matrix. Several methods are available to estimate the genetic relatedness present in isolated cohorts such as HELIC-MANOLIS[26]. We compared methods proposed in GEMMA[27], EMMAX[28], KING[29] and PLINK[30]., and found that the kinship coefficients reported by each method were highly correlated, but on a different scale from each other (Supplementary Fig. 9). For consistency with previous studies performed on the same samples, we calculated a genetic relatedness matrix using GEMMA[27] after filtering for MAF < 0.05, missingness <1% and LD-based pruning. In addition, MONSTER requires self-kinship coefficients on the

diagonal of the relatedness matrix, which we calculated using the $\widehat{F_1}$ metric from PLINK 1.9. The matrix was then converted to the long format using the reshape2 R package.

Association testing. Burden testing was performed using MONSTER[31], a method that extends the SKAT-O[32] model to account for relatedness and/or structure present in cohorts such as population isolates when testing for association. We ran burden testing across all genes defined in GENCODE v25 using 10 different conditions, i.e., combinations of regions of interest (coding regions only, coding and regulatory regions and regulatory regions only), variant filters (inclusion criteria based on severity of predicted consequence) and weighting schemes (Supplementary Table 1). QQ-plots for all testing conditions and traits are presented in Supplementary Fig. 10.

First, we extracted exonic coordinates for all protein-coding genes, which defines the region of interest for strictly exonic variants. These regions of interest were used in combination with 5 different variant filtering and weighting schemes. First, we included only variants predicted as high-confidence (HC) loss-of-function (LoF) by LOFTEE[8] that reside in the exons of protein-coding genes (Supplementary Table 1: LOFTEE HC). As only 460 variants in 85 genes passed this inclusion criterion, we performed an additional analysis including 8,570 low-confidence (LC) loss-of-function variants spread across 1,727 genes (Supplementary Table 1: LOFTEE LC). Stop-gained and frameshift mutations were the largest contributors to both the LC and HC sets. However, the LC set also includes a large number of splice donor and splice acceptor variants (Supplementary Fig. 4). We further performed an analysis with more relaxed inclusion criteria, including all exonic variants for which the Ensembl most severe consequence was more damaging than missense as predicted by the Variant Effect Predictor[33] (Supplementary Table 1: Exon severe). We also employed Combined Annotation Dependent Depletion (CADD)[34] scores, either to weigh all exonic variants (Supplementary Table 1: Exon CADD) or to filter out variants with CADD scores below the genome-wide median (Supplementary Table 1: Exon CADD median). Finally, we extended exon boundaries as defined above with 50 base pairs either side, to account for cases where potentially damaging variants occur on the edges of exons, as has been shown to happen for previously identified rare variant burdens[5]. These regions of interest were used in combination with one variant weighting scheme only (Exon + 50 CADD).

We extracted regulatory regions (promoters, enhancers and transcription-factor binding sites) from Ensembl build 84[12]. We assigned regulatory regions to genes if they directly overlapped or if the regulatory region overlapped with an eQTL for the gene based on the GTEx database[35]. If an eQTL was reported for several genes, overlapping variants were assigned to all of them. We did not take tissue specificity into account. For selecting variants, we either used the coordinates of the regulatory features alone, or regulatory features plus the extended exons. We used Eigen, an aggregate score that combines information from multiple regulatory annotation tracks[36], to weigh variants in all tests that include regulatory variants. In addition to raw Eigen scores, the authors also proposed EigenPC, a score derived from the first eigenvector of the correlation matrix of annotations. Both scores were available as is, or transformed using Phred-scaling, which maps a distribution's support to] 0,+∞[, thereby guaranteeing inclusion and relative up-weighting of all variants. In the regulatory regions plus exon analyses we used both the raw Eigen scores, shifted by 1 unit to the right, with negative scores set to $0 + \epsilon$ (Supplementary Table 1: Exon and regulatory Eigen), and the Phred-transformed Eigen and EigenPC scores (Supplementary Table 1: Exon and regulatory EigenPhred and EigenPCPhred). This transformation was a technical requirement as MONSTER could only read weights belonging to]0,+∞[. In the analyses containing the regulatory regions only, variants were weighted using the Phred-scaled Eigen scores (Supplementary Table 1: regulatory only EigenPhred) only.

Finally, we applied a MAF threshold of 0.05, a missingness threshold of 1% and a Hardy–Weinberg filter using a mid-p adjusted *P*-value[37] threshold of 1.0×10^{-5} to all variants prior to testing. We only performed a test if at least two SNVs passed the inclusion criteria for a given condition.

Establishing the significance threshold. We calculated $\alpha_{\text{eff}} = \frac{0.05}{N \times n_{\text{cond}} \times M}$, where N is the number of genes tested, n_{cond} is the effective number of inclusion and weighting criteria tested and M=6 is the number of traits. For n_{cond}, we plotted the correlation matrix of z-scores for all 10 analyses, and determined that the analyses using similar region definitions (exonic loss-of-function, exonic, exonic and regulatory variants) cluster together, reducing the effective number of analyses to 3 (Supplementary Fig. 7). Although $N = 18{,}997$ protein-coding genes are available in GENCODE V25, not all genes were tested in every condition. For example, for many genes only one variant might pass inclusion criteria in a high-confidence loss-of-function run, thereby excluding those genes from the analysis. A summary of the number of genes included in every analysis is presented in Supplementary Table 5. On average, $N = 13{,}854$ genes are included, hence we define study-wide significance at $P = 2.0 \times 10^{-7}$.

Burden prioritisation and novelty. We applied stringent checks to test the validity of rare variant burden association signals. Every suggestively associated burden (arbitrarily defined as $P \leq 5 \times 10^{-5}$) was conditioned on the genotypes of the variant included in the burden set with the lowest single-point *P*-value. If the *P*-value dropped more than two orders of magnitude below the suggestive significance threshold (i.e., $P \leq 5.0 \times 10^{-3}$), the burden was excluded from downstream analyses. We examined burden signals using the plotburden software (https://github.com/wtsi-team144/plotburden) to assess variant functionality, single-point association *P*-values, LD structure, as well as prior associations in the region. When a prior association was found in the region, we considered a signal known when the *P*-value dropped below $P = 1.0 \times 10^{-4}$ when conditioning on the genotypes of the existing signal. We examined rare variant burden associations with suggestive significance ($P < 5 \times 10^{-5}$) across the six traits under investigation, and do not find evidence of further rare variant signals at established loci.

Replication. The INTERVAL randomised controlled trial is a large-scale study focusing on healthy blood donors[10]. Sequencing, variant calling and quality control was performed for 3762 INTERVAL participants using the same protocol and pipeline as for the MANOLIS sequences. 38 samples were excluded on the basis of ethnicity, excessive relatedness ($\hat{\pi} > 0.125$), excess heterozygosity and contamination. VQSR thresholds of 99 and 90% for SNVs and INDELs, respectively, were applied to variant calls. Gamma-glutamyltransferase and adiponectin levels were not available in the INTERVAL replication cohort.

Selection analyses. For the selection analyses we used the haplotype-based iHS statistic[16]. We used this statistic because we were mainly interested in recent or ongoing selection, i.e., selective sweeps where the advantageous allele has not yet reached a high frequency (>80%), and iHS has been shown to be more powerful than other commonly used statistics like Tajima's D[38] and XP-EHH[39] for detecting such sweeps[16,40]. Briefly, the iHS value of an SNV becomes elevated when one of the alleles of that SNV reside on haplotypes that are longer than expected under neutrality, given the frequency of the allele. This is considered a signature of positive selection because positive selection will cause haplotypes carrying an advantageous allele to increase in frequency faster than if the allele had been neutral which leaves less time for recombination to shorten them.

To investigate if any of the four genes with study-wide significant burden association signals have undergone recent or ongoing positive selection, we calculated the fraction of SNVs with |iHS|>2 for these four genes and assessed if these fractions were elevated by comparing them to the empirical distribution for all genes. We focused on the fraction of SNVs with |iHS|>2, because previous studies have compared different methods of summarising iHS for a genomic region of interest, and found that the fraction of SNVs |iHS|>2 is often the most powerful iHS summary for detecting selection[16,40].

The primary data used for the selection analyses is the MANOLIS genotype data described above, which we phased using Beagle v.4.1[41]. We also used the ancestral allele annotations for each site in hg38 from ENSEMBL, and the recombination map from UCSC, which was built on hg37 and lifted-over to hg38. From this map, we excluded 3140 sites to achieve a recombination map with monotonically increasing cM positions. Linear interpolation was subsequently used to produce cM positions for all sites not found on the map. For quality control, we combined the MANOLIS genotype data with genotype data from the 1000 Genomes phase 3[42].

Because close relatives can complicate and potentially bias analyses of signals of selection, we removed close relatives within the MANOLIS data after phasing, as well as one admixed individual. We used the same criteria as in a previous study of this population[4], i.e., we used the --genome option in PLINK 1.9 to estimate PI-HAT and randomly excluded one individual from each pair of individuals with $\hat{\pi}$>0.2. These exclusions left 810 unrelated individuals from MANOLIS, on which we based the selection analysis.

We restricted our iHS analysis to known, common SNVs with ancestral allele annotations. Specifically, we excluded sites: not on the autosomes, with more than 2 alleles, with alleles that were not length 1 (INDEL-like), with MAF < 0.05, without ancestral allele annotations, with HWE midp-value < 1×10^{-30}, not present in 1000 Genomes phase 3 vcf files, or that were outside a mappable region of the hg38 reference genome, defined as a GEM 100-mer score below 0.8[43]. These filtering steps resulted in 5,126,987 SNVs as input for iHS calculations.

The iHS statistic was calculated with the hapbin program[44], using default parameters. The raw iHS statistic is sensitive to allele frequency, so SNVs were subsequently binned by derived allele count (82 equally-spaced bins) and the iHS statistic was normalised within each bin to have a mean of zero and a standard deviation of one, as suggested in[16]. Finally, we examine the absolute value of the normalised iHS statistic to capture selection signals associated with both derived and ancestral alleles. Due to edge effects at chromosome ends and other gaps, we examine iHS values for 5,116,861 SNVs (99.8% of input sites).

For each gene, we considered four distinct ways to define the genomic region representing the gene: (1) sites within exons, (2) sites within exons extended by 50 bp or in regulatory elements, (3) sites within the region spanned by connecting all exons, and (4) sites within the region spanned by connecting all exons extended by 50 bp and regulatory elements. For each gene and each of the four genic region definitions, we extracted SNVs with an iHS value and calculated the fraction of SNVs with normalised |iHS| above 2. When interpreting our results, we mainly focused on the results for the most inclusive definition, definition 4, as selection signatures tend to span fairly large genomic regions, but included the other

definitions to be able to assess if this choice of definition markedly affected our results.

For each lead gene with a rare variant study-wide significant burden signal (*APOC3, UGT1A9, ADIPOQ*, and *FAM189B*), we compared its fraction of SNVs with |iHS|>2 to all other genes with at least 1 iHS value-bearing SNV, using each of the four different gene region definitions. Each comparison was quantified by the percentile of genes with a higher fraction of SNVs with |iHS|>2. *FAM189B* was the only of the four burden genes with a fraction |iHS|>2 above zero. For this gene, we also performed a comparison to the subset of genes with a similar number of SNVs with iHS values as *FAM189B* (defined as +/− 10% of the number of SNVs with iHS in *FAM189B*) to ensure the varying number of SNVs in the genes we compared *FAM189B* to did not drastically affect the percentiles. Note that with the gene definitions used some SNVs will be included in several genes and thus the data points in the empirical distribution used for comparison are not entirely independent.

F_{ST} between two populations is a measure of population differentiation and is expected to increase in a region harbouring an allele which has been under positive selection mainly in one of the two populations.

To further investigate the *FAM189B* gene we calculated F_{ST} between the MANOLIS cohort and the European 1000 genomes CEU population for this gene and compared it to that of all other genes. The comparison was performed like for the iHS values, i.e., using quantiles and by performing several different comparisons to check for robustness of the results. For this analysis, we used the same genetic data from MANOLIS as for the iHS analyses combined with data from the 1000 genomes CEU population sample, except we did not filter away SNVs with MAF < 0.05, without ancestral allele annotations or with HWE midp-value $< 1 \times 10^{-30}$. Following published recommendations[45], all F_{ST} estimates were performed using the Hudson estimator and per SNP estimates were combined using the ratio of the average numerator and the average denominator (also referred to as weighted mean F_{ST}). However, we note that similar results were obtained using the Weir and Cockerham F_{ST} estimator[46].

Code availability. MUMMY, the script used to run burden tests genome wide using MONSTER, is available at https://github.com/wtsi-team144/burden_testing. The plotburden script, which builds interactive visualisations of burden signals, is available at https://github.com/wtsi-team144/plotburden.

Data availability

Sequencing data are available at the European Genome-Phenome Archive under accession numbers EGAS00001001207 for MANOLIS, EGAS00001000988 for TEENAGE, and EGAS00001001355, EGAS00001002461, and EGAS00001002787 for INTERVAL.

Received: 16 March 2018 Accepted: 8 October 2018
Published online: 07 November 2018

References

1. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).
2. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
3. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief. Funct. Genom.* **13**, 371–377 (2014).
4. Panoutsopoulou, K. et al. Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat. Commun.* **5**, 5345 (2014).
5. Gilly, A. et al. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum. Mol. Genet.* **25**, 2360–2365 (2016).
6. Farmaki, A. E. et al. The mountainous Cretan dietary patterns and their relationship with cardiovascular risk factors: the Hellenic Isolated Cohorts MANOLIS study. *Public Health Nutr.* **20**, 1063–1074 (2017).
7. Xue, Y. et al. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat. Commun.* **8**, 15927 (2017).
8. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
9. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
10. Moore, C. et al. The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
11. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
12. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
13. TG and HDL Working Group of the Exome Sequencing Project, N.H.L. et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* **371**, 22–31 (2014).
14. van Es, H. H. et al. Assignment of the human UDP glucuronosyltransferase gene (UGT1A1) to chromosome region 2q37. *Cytogenet. Cell Genet.* **63**, 114–116 (1993).
15. Sanna, S. et al. Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum. Mol. Genet.* **18**, 2711–2718 (2009).
16. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
17. Zhang, H. et al. Ectopic overexpression of COTE1 promotes cellular invasion of hepatocellular carcinoma. *Asian Pac. J. Cancer Prev.* **13**, 5799–5804 (2012).
18. Kallin, A. et al. SREBP-1 regulates the expression of heme oxygenase 1 and the phosphatidylinositol-3 kinase regulatory subunit p55 gamma. *J. Lipid Res.* **48**, 1628–1636 (2007).
19. Tachmazidou, I. et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4**, 2872 (2013).
20. Pollin, T. I. et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008).
21. Li, M. J. et al. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res.* **42**, D910–D916 (2014).
22. Gilly, A. et al. Very low depth whole genome sequencing in complex trait association studies. *bioRxiv* (2017).
23. Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).
24. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
25. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
26. Eu-Ahsunthornwattana, J. et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet.* **10**, e1004445 (2014).
27. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
28. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
29. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
30. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
31. Jiang, D. & McPeek, M. S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet. Epidemiol.* **38**, 10–20 (2014).
32. Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
33. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
34. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
35. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
36. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
37. Graffelman, J. & Moreno, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat. Appl. Genet. Mol. Biol.* **12**, 433–448 (2013).
38. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
39. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
40. Pickrell, J. K. et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
41. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
42. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
43. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
44. Maclean, C. A., Chue Hong, N. P. & Prendergast, J. G. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. *Mol. Biol. Evol.* **32**, 3027–3029 (2015).

45. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
46. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

Acknowledgements

HELIC-MANOLIS study: We thank the residents of the Mylopotamos villages for taking part. The MANOLIS study is dedicated to the memory of Manolis Giannakakis, 1978–2010. This work was funded by the Wellcome Trust [098051] and the European Research Council [ERC-2011-StG 280559-SEPI]. INTERVAL study: Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was funded by the National Institute of Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk/) and the NIHR Cambridge Biomedical Research Centre (www.cambridge-brc.org.uk). The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre. This report is independent research by the National Institute for Health Research. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. This work was undertaken by Cambridge who received funding from the NHSBT; the views expressed in this publication are those of the authors and not necessarily those of the NHSBT. TEENAGE study: The TEENAGE study has been supported by the Wellcome Trust (098051), European Union (European Social Fund—ESF) and Greek national funds through the Education and Lifelong Learning Operational Program of the National Strategic Reference Framework (NSRF)—Research Funding Program: Heracleitus II, Investing in knowledge society through the European Social Fund. The GATK3 program was made available through the generosity of the Medical and Population Genetics program at the Broad Institute, Inc. We acknowledge Giuseppe Matullo's contribution as EC's PhD supervisor.

Author contribution

Sample collection and phenotyping: A.E.F., I.N., E.T., J.D., G.D., E.Z. Sequencing Quality Control: A.G., D.S., K.H., E.C. Study design: A.G., D.S., K.K., T.B., E.V.R.A., E.Z. Association analyses: A.G., D.S., L.S. Software development: A.G., D.S. Bioinformatics: A.G., D.S. Selection analysis: R.W., I.M. Phenotype data management: K.K., G.M., B.K., N.W.R., A.B. Downsampling/depth analysis: A.G., M.P. Replication cohort analyses: A.G., KousikK, K.W. Manuscript writing: A.G., R.W., I.M., I.B., E.Z. Project supervision: E.Z.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07070-8.

Competing interests: The authors declare no competing interests.

nature COMMUNICATIONS

https://doi.org/10.1038/s41467-018-07730-9 OPEN

Author Correction: Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits

Arthur Gilly[1], Daniel Suveges[1], Karoline Kuchenbaecker[1,2,3], Martin Pollard[1,4], Lorraine Southam[1,5], Konstantinos Hatzikotoulas[1,6], Aliki-Eleni Farmaki[7,8], Thea Bjornland[9], Ryan Waples[10], Emil V.R. Appel[11], Elisabetta Casalone[12], Giorgio Melloni[13], Britt Kilian[1], Nigel W. Rayner[1,5,14], Ioanna Ntalla[15], Kousik Kundu[1,16], Klaudia Walter[1], John Danesh[1,17,18], Adam Butterworth[17,18,19], Inês Barroso[1], Emmanouil Tsafantakis[20], George Dedoussis[8], Ida Moltke[10] & Eleftheria Zeggini[1,6]

Correction to: *Nature Commununications*; https://doi.org/10.1038/s41467-018-07070-8, published online 7 Nov 2018

The original version of this Article contained an error in Fig. 2. In panel a, the two legend items "rare" and "common" were inadvertently swapped. This has been corrected in both the PDF and HTML versions of the Article.

Published online: 19 December 2018

[1] Department of Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom. [2] Division of Psychiatry, University College of London, London W1T 7NF, United Kingdom. [3] UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom. [4] Department of Medicine, Addenbrooke's Hospital, University of Cambridge, Hills Road, Cambridge CB2 0QQ, United Kingdom. [5] Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom. [6] Institute of Translational Genomics, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg D-85764, Germany. [7] Department of Health Sciences, College of Life Sciences, University of Leicester, Leicester LE1 6TP, United Kingdom. [8] Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Athens 176-71, Greece. [9] Department of Mathematical Sciences, Norwegian Institute of Science and Technology, Trondheim 7491, Norway. [10] Department of Biology, The Bioinformatics Centre, University of Copenhagen, 2200 Copenhagen, Denmark. [11] Section for Metabolic Genetics, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen 2200, Denmark. [12] Human Genetics Foundation, University of Torino, Torino IT-10126, Italy. [13] Department of Biomedical Informatics, Harvard Medical School, Boston 02115, MA, USA. [14] Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Old Road, Headington, Oxford OX3 7LE, United Kingdom. [15] William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, United Kingdom. [16] Department of Haematology, Cambridge Biomedical Campus, University of Cambridge, Long Road, Cambridge CB2 0PT, United Kingdom. [17] The National Institute for Health Research Blood and Transplant Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, University of Cambridge, Cambridge CB1 8RN, United Kingdom. [18] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Wort's Causeway, University of Cambridge, Strangeways Research Laboratory, Cambridge CB1 8RN, United Kingdom. [19] British Heart Foundation Centre of Excellence, Division of Cardiovascular Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, United Kingdom. [20] Anogia Medical Centre, 740 51 Anogia, Greece. Correspondence and requests for materials should be addressed to E.Z. (email: Eleftheria@sanger.ac.uk)

"Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits"

Gilly *et al.*

SUPPLEMENTARY INFORMATION

Supplementary Tables

Supplementary Table 1
Region definition, variant selection and weighting systems used to define testing conditions for burden analysis.

Burden analysis condition	Weighting system	Criterion	Exons	Regulatory Regions
LOFTEE HC	none	Predicted LoF by LOFTEE with high confidence	yes	no
LOFTEE LC	none	Predicted LoF by LOFTEE, both high and low confidence	yes	no
Exon severe	none	Ensembl most severe consequence more severe than missense	yes	no
Exon CADD	CADD	none	yes	no
Exon CADD median	CADD	CADD>5.851	yes	no
Exon+50 CADD	CADD	none	yes extended by 50bp	no
Exon+Regulatory Eigen	Eigen (raw score + 1)	Eigen>0	yes extended by 50bp	yes
Exon+Regulatory EigenPhred	Phred-transformed Eigen	none	yes extended by 50bp	yes
Exon+Regulatory EigenPCPhred	Phred-transformed EigenPC	none	yes extended by 50bp	yes
Regulatory only EigenPhred	Phred-transformed Eigen	none	no	yes

Supplementary Table 2
Burden test *P*-values for adiponectin levels in the *ADIPOQ* gene, conditioned on known adiponectin and diabetes-associated variants.

rsID	position (GRCh38) on chromosome 3	previous association	conditioned burden *P*-value
rs16861329	186948673	type 2 diabetes	4.76E-08
rs17366568	186852664	adiponectin levels	3.79E-08
rs182052	186842993	adiponectin levels	4.79E-08
rs822387	186838248	adiponectin levels, with and without BMI adjustment	2.56E-07
rs864265	186836503	adiponectin levels	5.56E-08
rs1648707	186833922	adiponectin levels	4.68E-08
rs10937273	186831906	adiponectin levels	5.24E-08
rs6810075	186830776	adiponectin levels	4.77E-08
rs266717	186812695	adiponectin levels	4.46E-08
rs266719	186783859	adiponectin levels	2.40E-07
rs822354	186762417	adiponectin levels	6.57E-08
rs74577862	186843903	adiponectin levels	1.2E-07
rs201813484	186841095	adiponectin levels	3.1E-07

Supplementary Table 3
Fraction of SNVs with |iHS|>2 in *APOC3*, *UGT1A9*, *ADIPOQ* and *FAM189B* compared to all other genes. For each gene its fraction of SNVs with |iHS|>2 is given in parenthesis and the percentile in the empirical distribution of these fractions for all genes using four different definitions of the genomic region representing the genes. We mainly considered the most inclusive definition (the bottommost), but included the others for comparison to assess robustness to this definition. For *FAM189B* the percentile is also given for the subset of genes with a similar gene length, defined as the number of SNVs with iHS values (rightmost column). A percentile of 80% means that 80% of values are less than or equal to the value.

Definition of burden testing condition	***APOC3* compared to all genes**	***UGT1A9* compared to all genes**	***ADIPOQ* compared to all genes**	***FAM189B* compared to all genes**	***FAM189B* compared only to genes with within +/- 10% of # SNVs in *FAM189B***
Exons only	41.0th percentile (0.00)	41.0th percentile (0.00)	41.0th percentile (0.00)	98.3th percentile (0.67)	97.4th percentile (0.67)
Exons extended by 50bp and regulatory elements	28.1th percentile (0.00)	28.1th percentile (0.00)	28.1th percentile (0.00)	97.4th percentile (0.42)	95.7th percentile (0.42)
Region spanning all exons	15.3th percentile (0.00)	15.3th percentile (0.00)	15.3th percentile (0.00)	96.7th percentile (0.32)	94.6th percentile (0.32)
Region spanning all exons extended by 50bp and regulatory elements	27.5th percentile (0.00)	27.5th percentile (0.00)	27.5th percentile (0.00)	95.6th percentile (0.33)	93.9th percentile (0.33)

Supplementary Table 4
Weighted mean F_{ST} in *FAM189B* compared to all other genes. The weighted mean F_{ST} for SNVs within *FAM189B* is given in parenthesis and the percentile of this value in the empirical distribution for all genes using four different definitions of the genomic region representing the genes (left column). In the right column, *FAM189B* is only compared to the subset of genes with a similar gene length, defined as the number of SNVs with F_{ST} values (rightmost column) within 10%.

Definition of gene	***FAM189B* compared to all genes**	***FAM189B* compared only to genes with within +/- 10% of # SNVs in *FAM189B***
Exons only	96.3th percentile (0.036)	96.8th percentile (0.036)
Exons extended by 50 bp and regulatory elements	99.1th percentile (0.050)	99.7th percentile (0.050)
Region spanning all exons	98.0th percentile (0.042)	97.5th percentile (0.042)
Region spanning all exons extended by 50 bp and regulatory elements	99.0th percentile (0.042)	100th percentile (0.045)

Supplementary Table 5
Number of genes with at least 2 SNVs for the different burden analysis conditions.

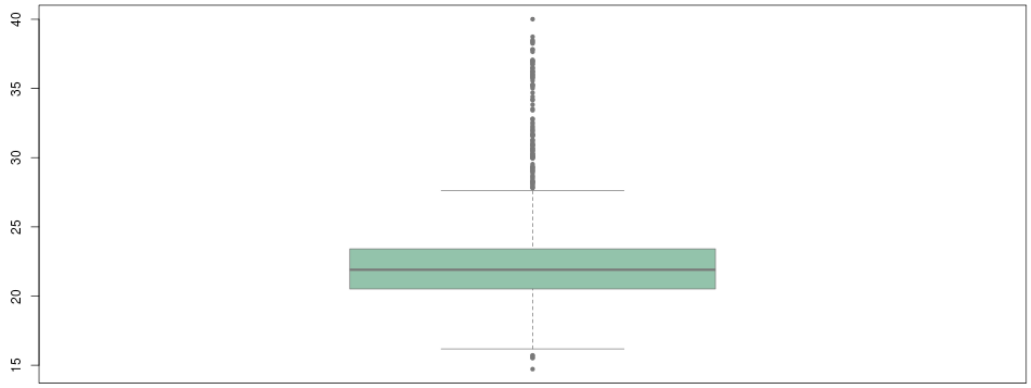
Analysis condition	Number of genes
GENCODE V25 (all protein-coding, not tested)	18,997
LOFTEE HC	85
LOFTEE LC	1,727
Exon severe	7,660
Exon CADD	18,428
Exon CADD median	18,138
Exon+50 CADD	18,551
Exon+Regulatory Eigen	18,961
Exon+Regulatory EigenPhred	18,660
Exon+Regulatory EigenPCPhred	18,722
Regulatory only EigenPhred	17,607

Supplementary Table 6
Broad functional categories are defined by grouping together several variant categories as defined by Ensembl VEP.

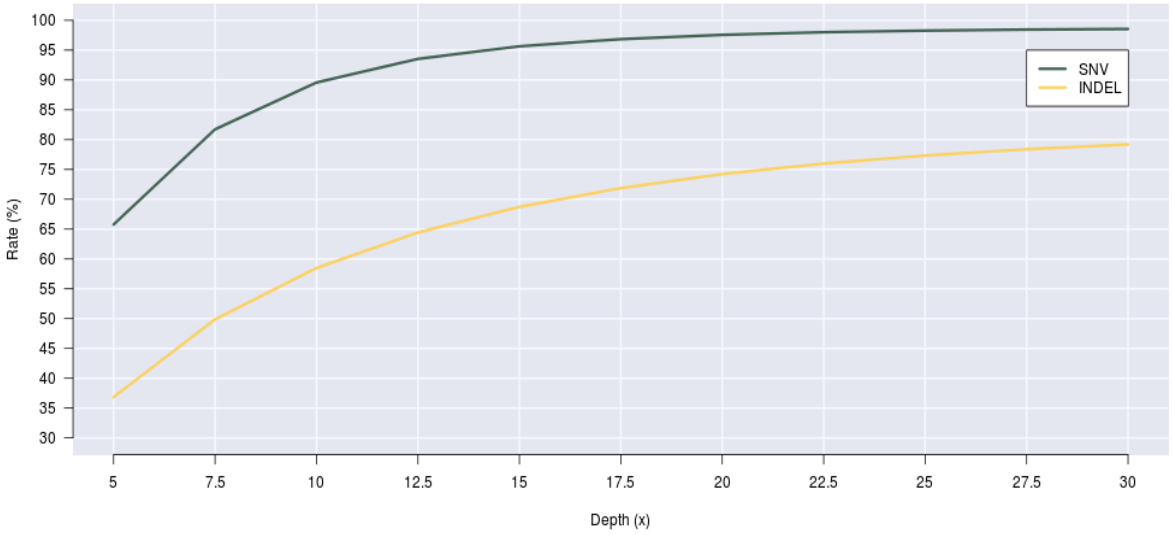
Attributed category	Ensembl VEP functional class
LoF (high)	*(See Supplementary Figure 4)*
LoF (low)	*(See Supplementary Figure 4)*
severe	start_lost stop_gained stop_lost frameshift_variant transcript_ablation
intergenic	intergenic_variant
Intronic	intron_variant
UTR	3_prime_UTR_variant 5_prime_UTR_variant
Up/Down stream	upstream_gene_variant downstream_gene_variant
splice variant	splice_donor_variant splice_acceptor_variant splice_region_variant
synonymous	synonymous_variant
other coding	coding_sequence_variant incomplete_terminal_codon_variant initiator_codon_variant missense_variant stop_retained_variant inframe_deletion inframe_insertion
other noncoding	nc_transcript_variant non_coding_transcript_exon_variant non_coding_exon_variant mature_miRNA_variant non_coding_transcript _variant
regulatory	regulatory_region_variant TF_binding_site_variant TFBS_ablation

Supplementary Figures

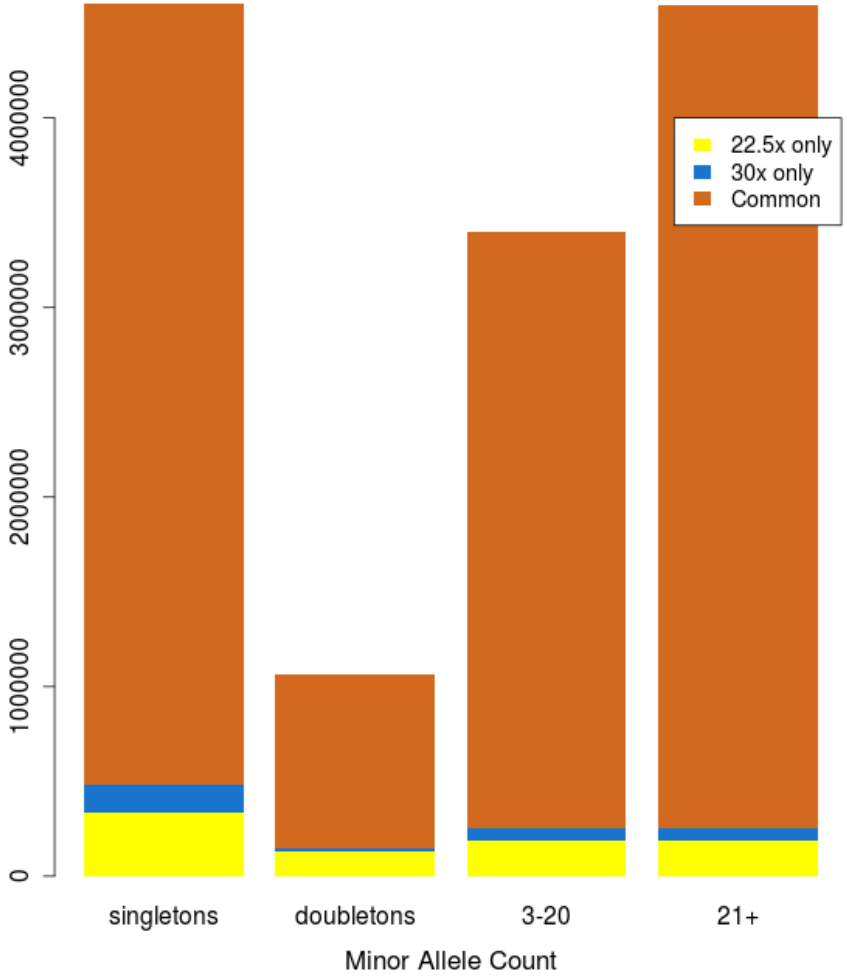
Supplementary Figure 1
Sequencing depth distribution of 1,457 MANOLIS samples. The box indicates quartiles and the centre line is the median. Whiskers extend to 1.5x the interquartile range. The mean is 22.5x and the median is 21.9x. Sequencing depths range from 14.7x to 40x.



Supplementary Figure 2
True positive rate for NA12878 at various sequencing depths. Compared to Genome in a Bottle 0.2 for SNVs and INDELs. Genome-wide, a single downsampling replicate was performed for each depth.

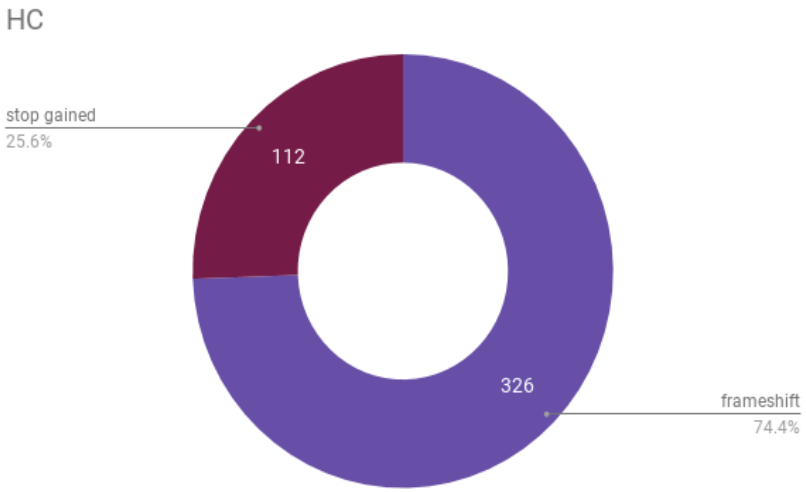


Supplementary Figure 3
Unique and shared SNVs between 30x and 22.5x depth whole genome sequencing.

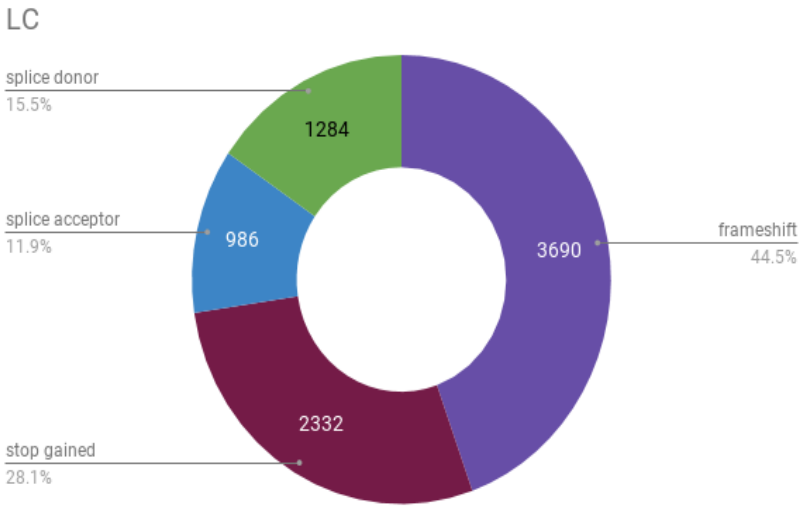


Supplementary Figure 4
Ensembl most severe consequence for loss-of-function variants. Breakdown is shown for variants predicted as loss-of-function with (a) high-confidence (HC) and (b) low-confidence (LC) by LOFTEE, along with genome-wide counts per Ensembl predicted consequence.

a.

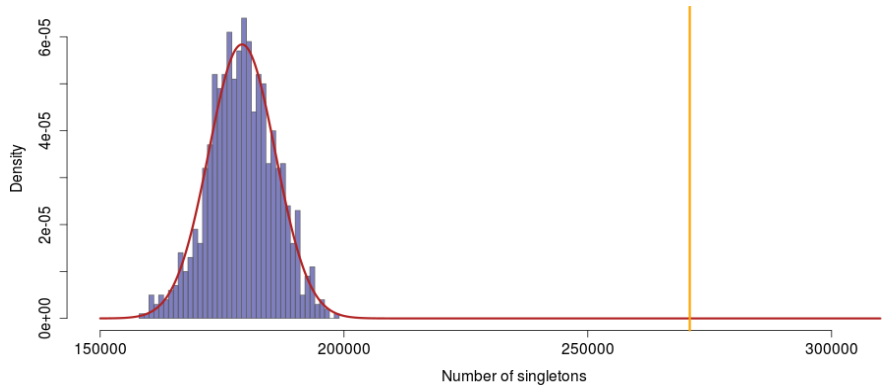


b.

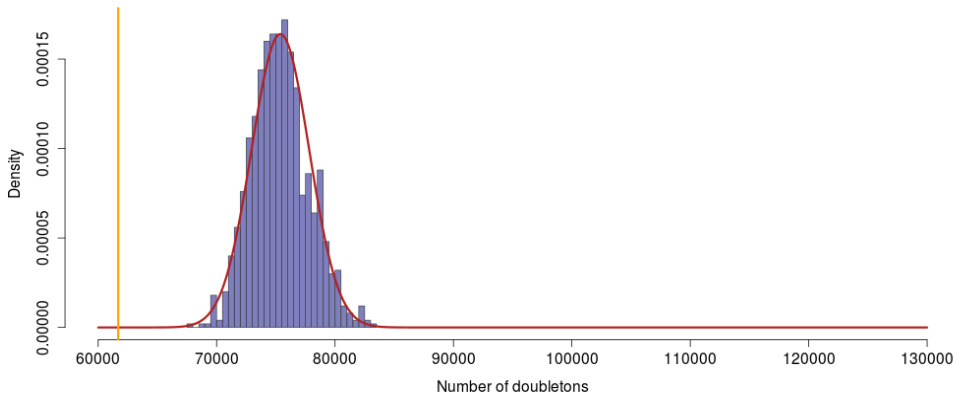


Supplementary Figure 5
Distributions of rare variant counts compared to cosmopolitan populations. (a) singleton and (b) doubleton counts in 1,000 draws of 100 MANOLIS samples (blue histograms). The vertical orange line indicates the observed count in 100 TEENAGE samples downsampled to 22.5x. Red lines are fitted normal distributions. (c) rare variant counts in MANOLIS (blue line) and INTERVAL (boxplots) in 500 draws of 1,482 samples from INTERVAL. Boxes extend from the 1st to the 3rd quartiles, whiskers extend to 1.5 times the interquartile range. Centre lines are the median.

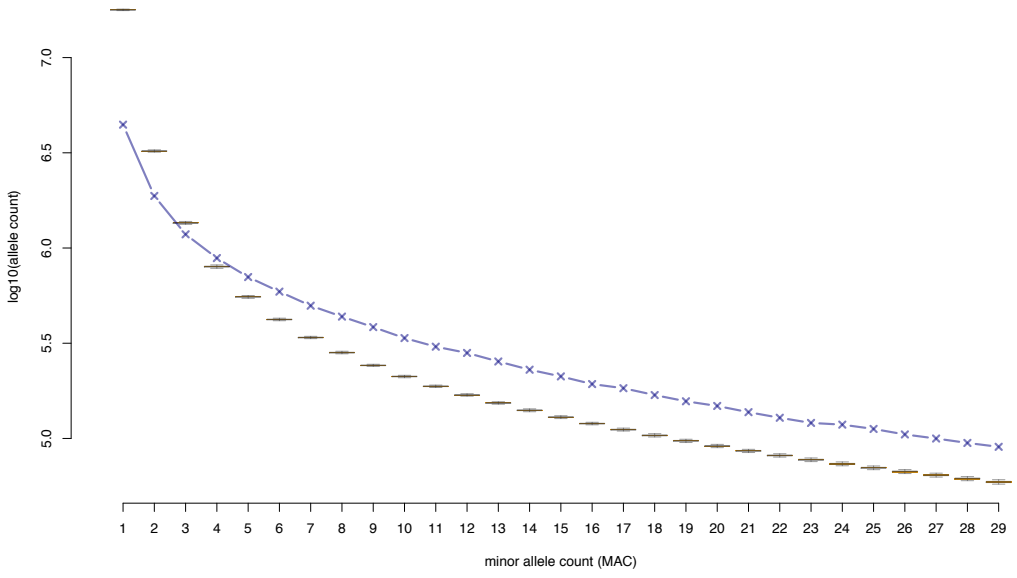
a.



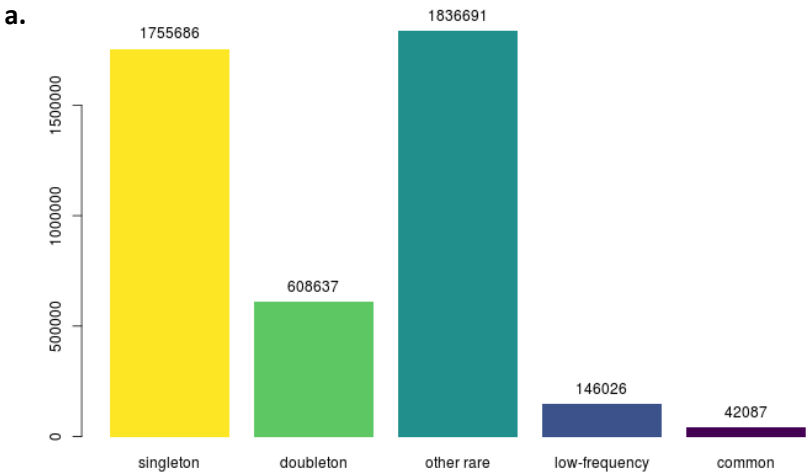
b.

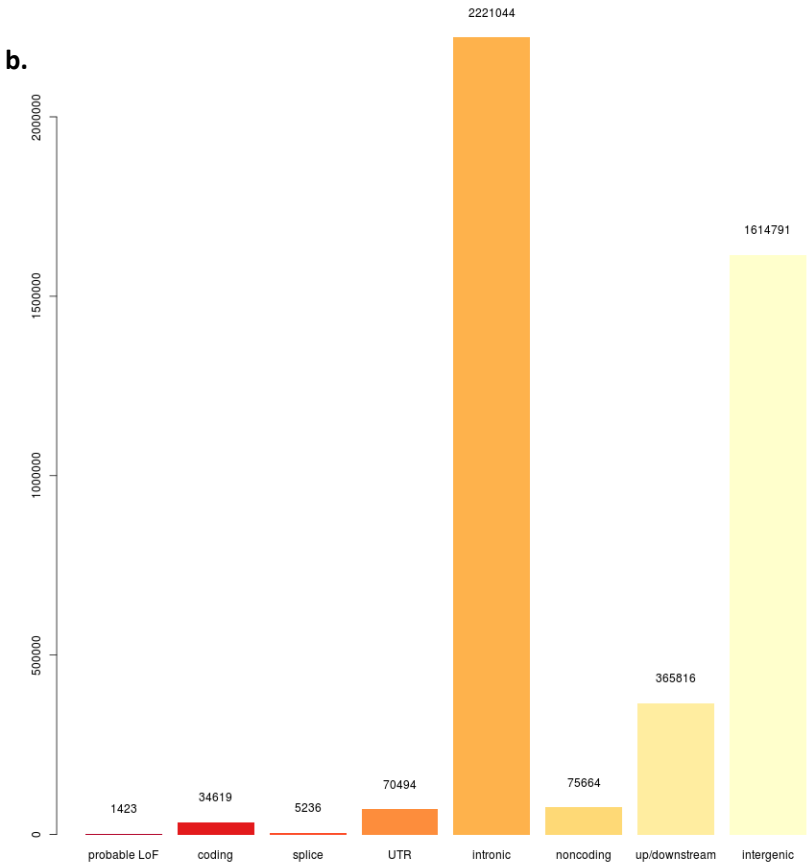


c.

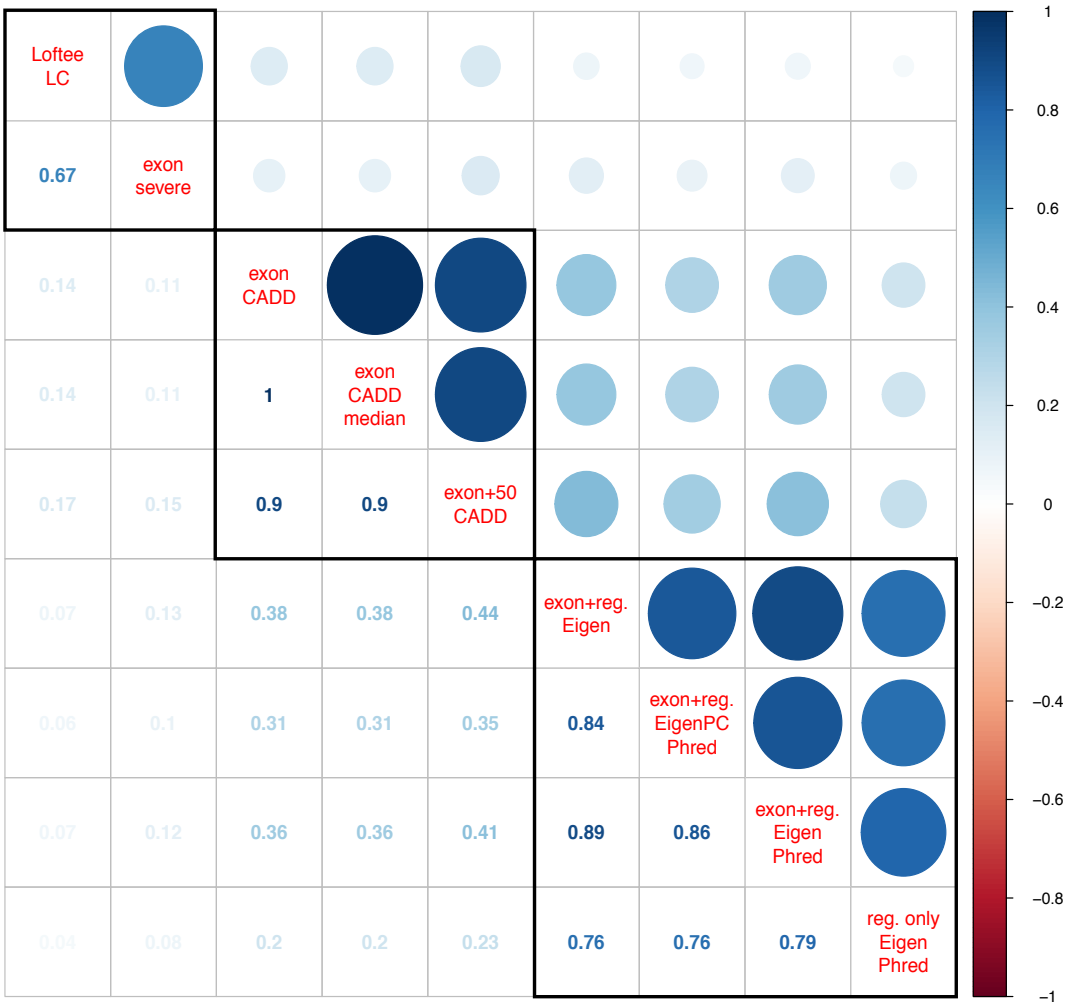


Supplementary Figure 6
Frequencies and functional annotations of all novel variants in MANOLIS. (a) depicts frequency bins and (b) variant consequences. Novelty is established by comparing variant location and alleles to Ensembl VEP annotation as well as gnomAD genomic variants lifted-over to build 38.



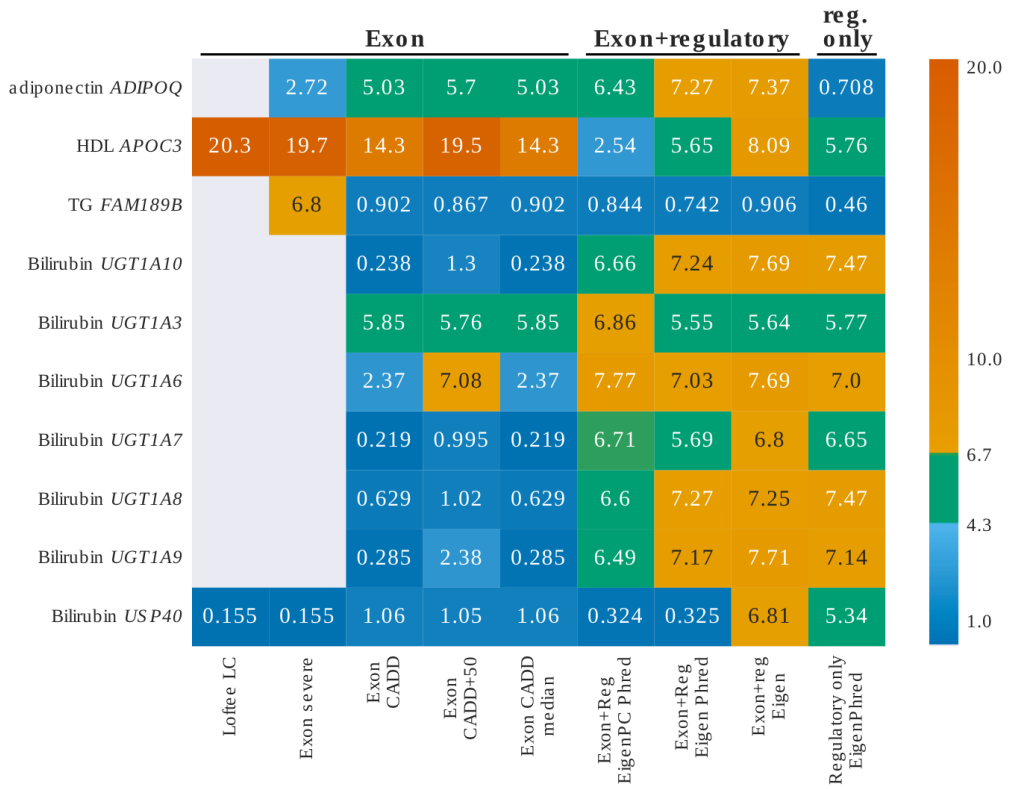


Supplementary Figure 7
Correlogram of z-scores arising from all evaluated burden testing scenarios. Quantities reported are Pearson's correlations of z-transformed *P*-values across all gene-trait pairs for all six tested traits. For each cell, *P*-values of all gene-trait pairs that were tested in both conditions are included. Clusters were generated using hierarchical clustering. The exon LoF HC scenario is not included in the correlogram due to the low number of genes containing more than one high-confidence loss-of-function variant (n=85).

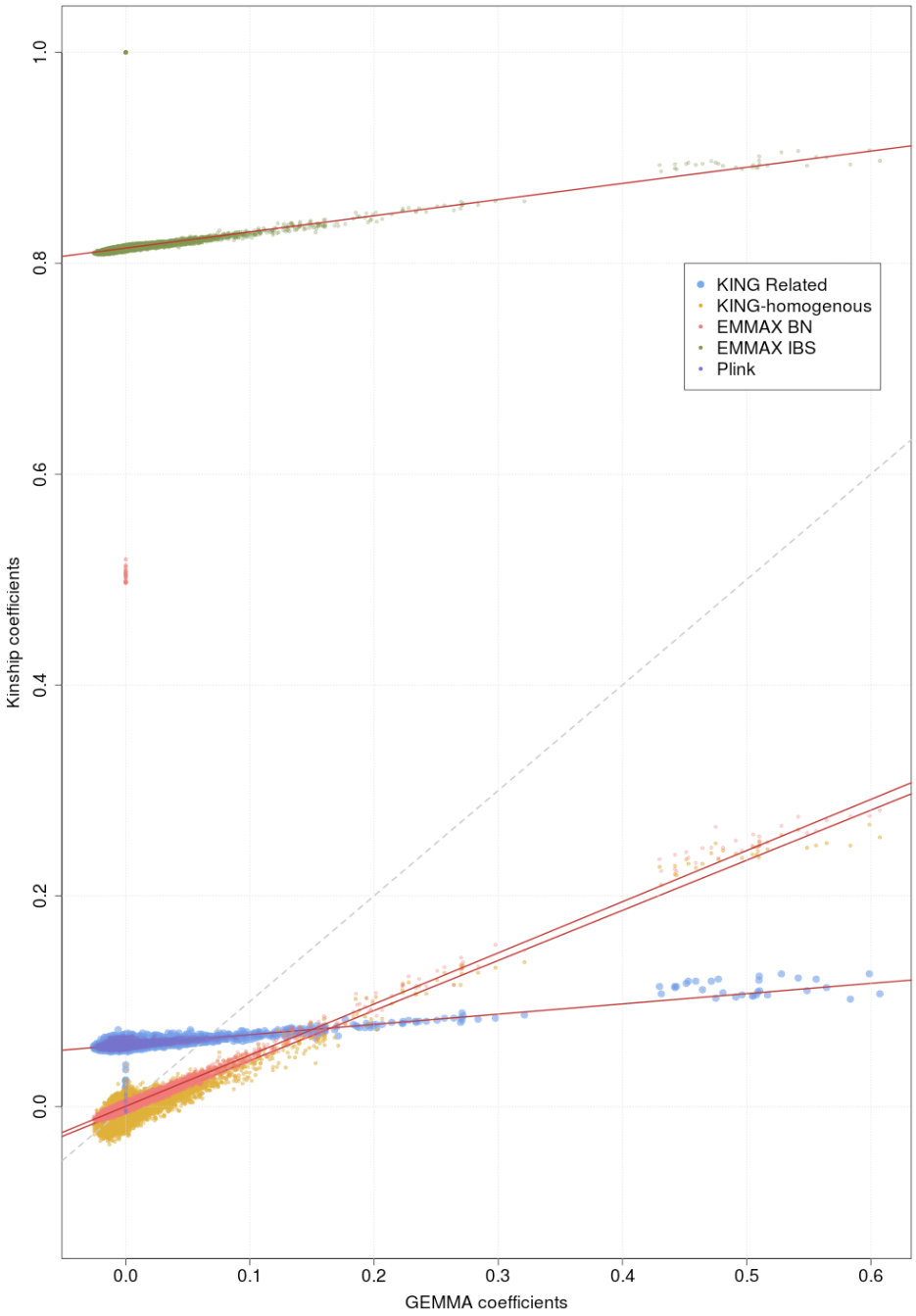


Supplementary Figure 8

Evidence for association for the 20 study-wide significant trait-gene pairs across all tested conditions. *P*-values are on the -$\log_{10}$ scale. Grey cells indicate that an insufficient number of variants passed the inclusion threshold. Dark orange denotes study-wide significance and turquoise green denotes suggestive association.



Supplementary Figure 9
Comparison of kinship coefficients produced by different methods. KING and EMMAX estimates on the y axis are compared to those produced by GEMMA[24] on the x axis. All kinship coefficients are calculated using the same dataset (MAF>5%, missingness<1%, LD-pruned). IBS coefficients (KING Related, EMMAX IBS and Plink) are both higher on average and less sensitive to increased relatedness than their Balding-Nichols counterparts (GEMMA, EMMAX BN and KING homogenous). Red lines represent OLS regression slopes.



Supplementary Figure 10
QQ-plots for all tested conditions across all analysed traits. The lambda values are displayed next to the condition name in the legend. lambda is calculated as

$$\lambda_{GC} = \frac{\text{median}(-log_{10}(p))}{-log_{10}(0.5)}$$