Weighted likelihood inference of genomic autozygosity patterns in dense genotype data

Alexandra Blant,[1,#] Michelle Kwong,[1,#] Zachary A. Szpiech,[2] and Trevor J. Pemberton[1,*]

[1]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada. [2]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California, USA.

[#]These authors contributed equally to this work

[*]Author to whom correspondence should be addressed.

Email addresses:

AB: umblant@myumanitoba.ca
MK: kwongmw@mcmaster.ca
ZAS: zachary.szpiech@ucsf.edu
TJP: Trevor.Pemberton@umanitoba.ca

Keywords: autozygosity; consanguinity; homozygosity; identity by descent; inbreeding; human populations

Abstract

Background: Genomic regions of autozygosity (ROA) arise when an individual is homozygous for haplotypes inherited identical-by-descent from ancestors shared by both parents. Over the past decade, they have gained importance for understanding evolutionary history and the genetic basis of complex diseases and traits. However, methods to detect ROA in dense genotype data have not evolved in step with advances in genome technology that now enable us to rapidly create large high-resolution genotype datasets, limiting our ability to investigate their constituent ROA patterns.

Results: We report a weighted likelihood approach for identifying ROA in dense genotype data that accounts for autocorrelation among genotyped positions and the possibilities of unobserved mutation and recombination events, and variability in the confidence of individual genotype calls in whole genome sequence (WGS) data. Forward-time genetic simulations under two demographic scenarios that reflect situations where inbreeding and its effect on fitness are of interest suggest this approach is better powered than existing state-of-the-art methods to detect ROA at marker densities consistent with WGS and popular microarray genotyping platforms used in human and non-human studies. Moreover, we present evidence that suggests this approach is able to distinguish ROA arising via consanguinity from ROA arising via endogamy. Using subsets of The 1000 Genomes Project Phase 3 data we show that, relative to WGS, intermediate and long ROA are captured robustly with popular microarray platforms, while detection of short ROA is more variable and improves with marker density. Worldwide ROA patterns inferred from WGS data are found to accord well with those previously reported on the basis of microarray genotype data. Finally, we highlight the potential of this approach to detect genomic regions enriched for autozygosity signals in one group relative to another based upon comparisons of per-individual autozygosity likelihoods instead of inferred ROA frequencies.

Conclusions: This weighted likelihood ROA detection approach can assist population- and disease-geneticists working with a wide variety of data types and species to explore ROA patterns and to identify genomic regions with differential ROA signals among groups, thereby advancing our understanding of evolutionary history and the role of recessive variation in phenotypic variation and disease.

Background

Genomic regions of autozygosity (ROA) reflect homozygosity for haplotypes inherited identical-by-descent (IBD) from an ancestor shared by both maternal and paternal lines. Common ROA are a source of genetic variation among individuals that can provide invaluable insight into how population history, such as bottlenecks and isolation, and "sociogenetic" factors, such as frequency of consanguineous marriage, influence genomic variation patterns. Population-genetic studies in worldwide human populations over the past decade have found ROA ranging in size from tens of kb to multiple Mb to be ubiquitous and frequent even in ostensibly outbred populations [1-28] and to have a non-uniform distribution across the genome [7,10,13,18] that is correlated with spatially variable genomic properties [2-4,18] creating autozygosity hotspots and coldspots [18]. ROA of different sizes have different continental patterns both with regards to their total lengths in individual genomes [12,18,22,24,26-28] and in their distribution across the genome [18] reflecting the distinct forces generating ROA of different lengths. Studies of ROA in the genomes of ancient hominins [29-31] and early Europeans [32] have provided unique insights into the mating patterns and effective population sizes of our early forbearers. In non-humans, ROA patterns have provided insights into the differential histories of woolly mammoth [33], great ape [34,35], cat [36], canid [37-42], and bird [43] populations, while in livestock breeds they have provided insights into their origins, relationships, and recent management [42,44-61] and the lasting effects of artificial section [58,61-73], as well as informed the design of ongoing breeding [74,75] and conservation [47,57,76] programs [77].

In contemporary human populations, increased risks for both monogenic [78-82] and complex [83-90] disorders as well as increased susceptibility to some infectious diseases [91-93] have been observed among individuals with higher levels of parental relatedness. While the association between parental relatedness and monogenic disease risk has been known for more than a century [94], observations with complex and infectious diseases potentially reflect elevated levels of autozygosity as a consequence of prescribed and unintentional inbreeding [95] that enrich individual genomes for deleterious variation carried in homozygous form [96,97]. Indeed, genomic autozygosity levels have been reported to influence a number of complex traits, including height and weight [98-101], cognitive ability [101-103], blood pressure [104-111], and cholesterol levels [111], as well as risk for complex diseases such as cancer [84,85,112-116], coronary heart disease [84,117-119], amyotrophic lateral sclerosis (ALS) [120], and mental disorders [121,122]. These observations are consistent with the view that variants with individually small effect sizes associated with complex traits and diseases are more likely to be rare than to be common [123-126], are more likely to be distributed abundantly rather than sparsely across the genome [9,127], and are more likely to be recessive than to be dominant [9,128]. Recent studies investigating ROA and human disease risk have identified both known and novel loci associated with standing height [129], rheumatoid arthritis [130], early-onset Parkinson's disease [131], Alzheimer's disease [132,133], ALS [120], schizophrenia [4,134], thyroid cancer [116], and Hodgkin lymphoma [115,135]. Thus, just as ROA sharing among affected individuals has facilitated our understanding of the genetic basis of monogenic disorders [136] in both inbred [137-140] and more outbred [141-143] families, it also represents a potentially powerful approach with which to further our understanding of the genetic etiology of complex disorders [144] of major public health concern worldwide.

In both population- and disease-genetic studies, ROA are frequently inferred from runs of homozygous genotypes (ROH) present in genome-wide single nucleotide polymorphism (SNP) data obtained using high-density microarray platforms [145]. A popular program for ROH identification is *PLINK* [146], which uses a sliding window framework to find stretches of contiguous homozygous genotypes spanning more than a certain number of SNPs and/or kb, allowing for a certain number of missing and/or heterozygous genotypes per window to account for possible genotyping errors. While a number of more advanced ROA identification approaches have been proposed [147,148], a recent comparison found the *PLINK* method to outperform these alternatives [149]. We recently proposed to detect ROA using a sliding-window framework and a logarithm-of-the-odds (LOD) score measure of autozygosity [1,150] that offers several key advantages over the *PLINK* method [18]. First, it is not reliant on fixed parameters for the number of heterozygous and missing genotypes when determining the autozygosity status of a window, instead incorporating an assumed genotyping error rate, making it more robust to missing data and genotyping errors. Second, it incorporates allele frequencies in the general population to provide a measure of the probability that a given window is homozygous by chance, allowing homozygous windows to be distinguished from autozygous windows. These important advances would be expected to provide greater sensitivity and specificity for the detection of ROA in high-density SNP genotype data, particularly in the presence of the higher and more variable genotype error rates in next-generation sequence (NGS) data [151,152].

A shortcoming of the LOD method is that correlations between SNPs within a window that occur as a consequence of linkage disequilibrium (LD) are ignored, leading to overestimation of the amount of information that is available in the data and potentially false-positive detection of autozygosity signals. In addition, the LOD method does not account for the possibility of recent recombination events onto very similar haplotype backgrounds that might give the appearance of autozygosity when paired with a non-recombined haplotype [153]. Such a scenario would, for example, arise when ROA are detected in microarray-based genotype data that comprises information at only a limited set of positions within a genomic interval and is therefore blind to unobserved genetic differences that make the apparently identical haplotypes distinct.

Here, we report an improved LOD-based ROA detection method that accounts for the non-independence between SNPs and the likelihoods of unobserved mutation and recombination events within a window. We compare the performance of this new method against the original LOD method as well as a newly reported method implemented in the *BCFtools* software package [154] in simulated genetic datasets. We then evaluate how ROA inference is influenced by the source and density of interrogated markers using the 26 worldwide human populations included in Phase 3 of The 1000 Genomes Project [155], considering the entire whole-genome sequence (WGS) dataset as well as subsets representing SNPs present in the exome and included on two commonly used Illumina BeadChips. We show population differences in genome-wide ROA patterns inferred from WGS data using our improved LOD-based method recapitulate those observed in our earlier BeadChip-based study that used the original LOD method [18]. Finally, we highlight the unique ability of our improved LOD-based method to identify genomic regions enriched for autozygosity signals in one group relative to another without first inferring ROA through the direct comparison of weighted LOD scores, finding nine regions that significantly differ in the strength of their autozygosity signals between apparent subgroups within the Asian Indian Gujarati, Punjabi, and Telugu populations. Our improved ROA detection method will assist population- and disease-geneticists working with a wide variety of data types and species to explore ROA patterns and to identify genomic regions with differential ROA signals, thereby facilitating our understanding of the role of recessive variants in phenotypic variation and disease.

Results

Weighted likelihood autozygosity estimator

We previously reported an ROA detection approach that was based on a number of earlier methods [1,150] in which a likelihood-based autozygosity estimator is applied in a sliding window framework where window size is defined as a fixed number of SNPs [18]. In this approach, within window w in individual i from population j, the LOD score of autozygosity is calculated across the K SNP markers within window w, where we observe genotype G_k at the k^{th} SNP that has state X_k, which equals 1 if the SNP is autozygous and 0 otherwise.

$$LOD(w,i) = \sum_{k=1}^{K} log_{10}\left(\frac{\Pr(G_k|X_k = 1)}{\Pr(G_k|X_k = 0)}\right) \qquad (1)$$

The per-SNP likelihoods of autozygosity and non-autozygosity are based on Hardy-Weinberg proportions (**Table 1**) and include population-specific allele frequencies and an assumed rate of genotyping errors and mutations ε. Missing genotypes are ignored in this algorithm; that is, they have a log-likelihood of zero.

The log-likelihood of autozygosity for homozygous SNPs is positive and decreases exponentially as a function of allele frequency (Additional File 1: **Figure S1A**). The log-likelihood of autozygosity for heterozygous SNPs is instead negative and equal to $\log_{10}(\varepsilon)$, thus acting as a penalty for the presence of heterozygous genotypes within a window.

To address the apparent shortcomings of the LOD score method, we developed a weighted LOD-based method ($wLOD$) that accounts for non-independence among SNPs and the probabilities of recombination and mutation within window w.

$$wLOD(w,i) = \sum_{k=1}^{K} \left[\begin{array}{r} log_{10}\left(\frac{\Pr(G_k|X_k=1)}{\Pr(G_k|X_k=0)}\right) \times \mathrm{Corr}(p_k,[p_1,p_K]) \\ \times \Pr(no\ recombination|[g_{k-1},g_k]) \\ \times \Pr(no\ mutation|\mu,[p_{k-1},p_k]) \end{array} \right] \tag{2}$$

Here, we adapt the approach of Chen *et al*. [156] to incorporate LD information into the $wLOD(w,i)$ estimator, weighting the log-likelihood of SNP k by the reciprocal of the sum of pairwise LD between SNP k and all other SNPs within window w calculated as

$$\mathrm{Corr}(p_k,[p_1,p_K]) = \frac{1}{\sum_{l=1}^{K} LD_{k,l}} \tag{3}$$

and bounded in the interval $[1/K,1]$. An intuitive explanation for this correction is that when a number of SNPs are highly correlated they provide redundant information. By weighting the log-likelihood for SNP k as a function of its correlation with all other SNPs within window w it contributes only the unique autozygosity information it possesses to w$LOD(w,i)$.

LD reflects historical recombination and mating patterns in a population and is largely insensitive to the effects of mating patterns within the last few generations that can, through recombination events onto very similar haplotype backgrounds, lead to false-positive autozygosity signals [153]. Thus, we also weight the log-likelihood of SNP k by the probability of no recombination events having occurred within the genomic interval bounded by SNP $k-1$ and SNP k in the last M generations, calculated based upon their genetic map position g (in Morgans) as previously described [10,157]

$$\Pr(no\ recombination|[g_{k-1},g_k]) = e^{-2M(g_k-g_{k-1})} \tag{4}$$

In a population-genetic context, M can be set based upon effective population size estimates and the probability that a pair of individuals share a common ancestor M generations in the past [158], while in a disease-genetic context M can instead be set based on known relationships between affected individuals.

Finally, we account for the potential presence of unobserved genetic differences within the genomic interval bounded by SNP $k-1$ and SNP k by weighting the log-likelihood of SNP k by the probability of no unobserved mutation events having occurred within the genomic interval in the last M generations, calculated based upon their physical map position p (in bp) and a per-base mutation rate μ using an approach adapted from MacLeod *et al*. [159]

$$\Pr(no\ mutation|\mu,[p_{k-1},p_k]) = e^{-2M\mu(p_k-p_{k-1})} \tag{5}$$

As evident in **Figure S1B** (Additional File 1), the recombination and mutation weightings reduce the log-likelihood of SNP k as a function of its distance from SNP $k-1$. It can also be seen that as M decreases the magnitude of the change in the weighting with increasing distance also decreases; thus, $wLOD$ scores in populations with small effective population sizes or in disease studies where affected individuals share a more recent common ancestor (smaller M) will be adjusted to a lesser extent than those with larger effective population sizes or where affected individuals share a more distant common ancestor (larger M).

Properties of the *wLOD* estimator

We investigated the properties of the $wLOD$ estimator using The 1000 Genomes Project Phase 3 dataset that provides phased genotypes for 84,801,880 genetic variants discovered using a low-coverage WGS approach in 2,436 unrelated individuals from 26 worldwide human populations (**Table 2**) [155]. To approximate a typical microarray-based SNP genotyping study, we first developed a subset of this dataset that contained 2,166,414 autosomal SNPs that are present on the popular Illumina HumanOmni2.5-8 BeadChip ("Omni2.5 dataset" henceforth). In all analyses, μ was set to 1.18×10^{-8} [160] and ε was set to 4.71×10^{-4}, the average rate of discordance across samples between genotypes in our Omni2.5 dataset and those obtained for 1,693 of the 2,436 individuals directly with the Illumina HumanOmni2.5 BeadChip [155]. Unless otherwise stated, M was set to seven, a conservative value broadly reflecting the average of effective population size estimates for populations included in The 1000 Genome Project [155,158,161]. Window size was varied in an arbitrary interval $[K_0, K_n]$ in which K is increased in 10 SNP increments (i.e. $K_n = K_0 + [10\times n]$).

The genome-wide distribution of $wLOD$ scores for all windows in the Omni2.5 dataset is bimodal and centered around 0 (**Figure 1A**), with $wLOD$ scores under the left-hand mode favoring the hypothesis of non-autozygosity, whereas those under the right-hand mode favor the autozygosity hypothesis. The area under the right-hand mode decreases as a function of window size as ROA are progressively covered by fewer but longer windows. In addition, while the location of the right-hand mode does not change appreciably with window size, there is a noticeable shift toward lower $wLOD$ scores in the left-hand mode with increasing window size, likely reflecting the larger number of heterozygous SNPs in non-autozygous compared with autozygous regions and their greater cumulative effect on $wLOD$ scores with increasing window size. This shift progressively increases the distance between the non-autozygous and autozygous modes until either the autozygous mode disappears (**Figure 1B**) or the intermodal distance begins to decrease instead (Additional File 1: **Figure S2**), both potentially reflecting the point above which window lengths exceed those of the majority of ROA in the sample set. In this scenario, as window size increases autozygous windows increasingly overlap non-autozygous regions flanking shorter ROA leading them to encompass greater numbers of heterozygotes within these non-autozygous regions, deflating their $wLOD$ scores. Whether the autozygous mode disappears or shifts toward lower $wLOD$ scores is likely determined by the abundance of ROA and their levels of support in the sample set: sets with fewer ROA and ROA with generally lower $wLOD$ scores lose their autozygous mode while those with large numbers and higher $wLOD$ scores have it shift toward the non-autozygous mode. Nevertheless, the location of the minimum between the two modes does shift subtly toward higher $wLOD$ scores with increasing window size, reflecting the expected increase in scores for autozygous windows as a function of the number of SNPs within the window. The periodicity observed in the genome-wide score distribution of the original LOD estimator [18] is absent with the $wLOD$ estimator, indicating that this property was a reflection of LD among SNPs within the window.

To evaluate how the improvements incorporated into the $wLOD$ estimator (equation 3) influence per-window scores as compared to the original LOD estimator (equation 1), we compared $wLOD$ and LOD scores in the Omni2.5 dataset with a window size of 150 SNPs (**Figure 2A**), the largest value that produced a clear bimodal $wLOD$ score distribution in all populations. Across populations, per-window $wLOD$ scores differed from their corresponding LOD scores by between -103.87 and 454.07 (**Figure 2B**) with the range and average of $wLOD$ and LOD score differences increasing as a function of a population's geographic distance from East Africa (ρ=0.8460 with P=5.029×10^{-6} and ρ=0.8846 with P=4.961×10^{-7}, respectively), reflecting increasing LD [162,163] and decreasing genetic diversity [95,164-167]—leading to larger inter-SNP distances—with distance from Africa. Among the six admixed populations included in Phase 3 of The 1000 Genomes Project, those of mixed African and European ancestry (ACB and ASW) had smaller ranges and averages of $wLOD$ and LOD score differences than those of mixed of Amerindian and European ancestry (CML, MXL, PUR, and PEL), consistent with the lower LD [168-170] and higher genetic diversity [167,171] of admixed African-European populations compared with Amerindian-European populations.

Across populations, 5.15–47.93% of all windows in the right-hand "autozygous" mode with the LOD estimator were present in the left-hand "non-autozygous" mode with the $wLOD$ estimator (**Figure 2C**) potentially reflecting false-positive autozygosity signals reported by the LOD estimator as a consequence of non-independence among homozygous SNPs that cumulatively give the mistaken impression of autozygosity. The proportion of windows was lowest in African populations and highest in most European populations, increasing incrementally through Central/South Asian and East Asian populations. This pattern can be explained by population differences in the location of the autozygous mode and its shift toward lower scores with the $wLOD$ estimator. Modal LOD and $wLOD$ scores in the autozygous mode are generally smallest and most similar in European populations and highest and least similar in African populations (Additional File 1: **Figure S3A**). Thus, for a given unit decrease in score between the LOD and $wLOD$ estimators, an autozygous LOD window has a greater chance of transitioning to the non-autozygous $wLOD$ mode in Europeans populations than in African populations. Consistent with this hypothesis, the magnitude of the difference between modal LOD and $wLOD$ scores in the autozygous mode and the location of the minima between the autozygous and non-autozygous modes is significantly negatively correlated with the proportion of autozygous LOD windows that transition to the non-autozygous $wLOD$ mode (r=-0.8654, P=1.156×10^{-8}; Additional File 1: **Figure S3B**).

In contrast, across populations only 0.055–5.015% of all windows in the non-autozygous mode with the LOD estimator were present in the autozygous mode with the $wLOD$ estimator (**Figure 2D**), potentially reflecting false-negative autozygosity signals reported by the LOD estimator as a consequence of heterozygotes in high LD with a larger number of homozygotes that, in one possibility, might reflect genotyping errors. The proportion of windows was highest in most African populations and lowest in most European populations, with broadly similar values observed in Central/South and East Asian populations. This pattern is the opposite of that observed with the putative false-positive windows above, and can also be explained by population differences in the location of the autozygous mode and its shift toward lower scores with the $wLOD$ estimator. The addition of a single heterozygote to an autozygous window in the European populations has a greater chance of transitioning it from the autozygous to non-autozygous mode than in the African populations since the autozygous mode is situated much closer to the minima between the two modes (Additional File 1: **Figure S3**).

Overall, the much larger numbers of windows transitioning from the autozygous to the non-autozygous mode than vice versa between the LOD and $wLOD$ estimators accords with the expectation that the LOD estimator frequently overestimates the amount of information available in the data leading it to falsely report autozygosity signals particularly in genomic regions with higher levels of LD, while it underestimates the amount of information much less frequently.

Evidence of separate endogamic and consanguinity autozygosity signals in Asian Indians

In four of the five Asian Indian populations─Gujarati (GIH), Telugu (ITU), Punjabi (PJL), and Sri Lankan Tamil (STU)─as well as in the East Asian Dai (CDX) population, as window size increased a third mode appeared in their $wLOD$ score distribution that divided the right-hand autozygous mode in two (**Figure 3A**). While an apparent third mode also appeared in the $wLOD$ score distribution of the Bengali (BEB) Asian Indian population, it was not as well defined as those of the other populations. As window size increased further, the area under both autozygous modes decreased until the left-hand autozygous mode disappeared followed sometime later by the right-hand autozygous mode. Notably, the distributions of all other populations in our dataset did not develop this third mode, and trimodality was not observed in the distribution of LOD scores for any of the 26 populations in the Omni2.5 dataset.

The appearance of a trimodal distribution in these six populations potentially reflects the effects of two distinct cultural processes that occur in India and among the Dai: consanguinity [172,173] and endogamy [174,175]─the restriction of marriages to within a predefined group of lineages or villages. In this scenario, the right-hand autozygous mode represents ROA due to consanguinity that are enriched for alleles rare in the general population that segregate within inbred families, while the left-hand autozygous mode represents ROA due to endogamy that are enriched for alleles present at low frequency in the general population that segregate within specific endogamic groups. Compatible with this hypothesis, the three populations with the strongest trimodal pattern (STU, ITU, and DAI) have higher reported frequencies of consanguinity (38.2% [173], 30.8% [173], and 21.3% [172]) than those with weaker trimodal patterns (BEB, 5.0% [173]; GIH, 4.9% [173]; PJL, 0.9% [173]). For example, the consanguinity-associated mode of the ITU is much larger than the endogamy-associated mode, while the reverse is true for the GIH (**Figure 3A**), consistent with consanguinity being the primary force generating ROA in the ITU while endogamy is the dominant force in the GIH. To the best of our knowledge, none of the other populations included in Phase 3 of The 1000 Genomes Project practise endogamy; consequently, we do not observe a separate endogamy-associated autozygous mode in their $wLOD$ score distributions.

If trimodal distributions are indeed a reflection of the $wLOD$ method being able to disentangle autozygosity signals arising from endogamy and consanguinity processes we would expect inferred ROA to be delineated predominantly by windows from only one of the two autozygous modes. Conversely, if the trimodal distribution is just an idiosyncrasy of the $wLOD$ estimator we would instead expect ROA to be delineated by a random mix of windows drawn from the two autozygous modes. To investigate how windows in the putative endogamy- and consanguinity-associated modes cluster to form inferred ROA, separately for each population exhibiting a clear trimodal $wLOD$ score distribution, we constructed ROA from windows with $wLOD$ scores above the minimum between the non-autozygous and left-most autozygous modes in their $wLOD$ score distribution [18]. Next, for each inferred ROA, we calculated the proportion of their underlying autozygous windows that had $wLOD$ scores within the right-most putative consanguinity-associated mode (i.e. above the minimum between the two autozygous modes).

Inferred ROA were found to frequently be delineated by windows drawn predominantly from one of the two autozygous modes (**Figure 3B**). A large well-defined peak is observed at low proportions, representing those ROA comprised of >90% of windows drawn from the left-hand endogamy-associated mode. A more diffuse peak is observed at higher proportions, representing those ROA comprised of >80% of windows drawn from the right-hand consanguinity-associated mode. The dispersed appearance of the peak representing putative consanguinity-associated ROA can be explained as a reflection of the fact that the two autozygous modes are not distinct. At the ends of ROA arising via consanguinity, the $wLOD$ scores of windows will naturally decrease as they increasingly span non-autozygous regions and overall support for autozygosity declines, leading them to increasingly fall within the endogamy-associated mode. Consequently, we would expect ROA arising via consanguinity to contain a small proportion of windows in the endogamy-associated mode, with the proportion varying based upon the overall strength of the autozygous signal (i.e. ROA conferring generally higher $wLOD$ scores will have lower proportions of windows in the endogamy-associated mode). Nevertheless, across populations, 68.9% (PJL) to 84.5% (CDX) of all ROA had >80% of their component windows drawn from a single autozygous mode.

Additional support for trimodality in the $wLOD$ score distribution reflecting distinct autozygosity signals arising from endogamy and consanguinity processes is provided by a comparison of how the proportion of windows drawn from the consanguinity-associated mode changes with ROA length (Additional File 1: **Figure S4**). Almost all ROA longer than 5 Mb are comprised predominantly of windows drawn from the consanguinity-associated mode (>90%), while proportions among ROA shorter than 5Mb are much more variable. This pattern is consistent with the expectation that ROA arising via consanguinity will in general be much longer than those arising via endogamy.

Overall, the properties of ROA constructed from the trimodal $wLOD$ score distributions present in the Asian Indian and East Asian Dai populations are compatible with the $wLOD$ method being capable of disentangling autozygosity signals that arise from different cultural processes at sufficiently large window sizes. However, further work in well-defined populations that practise both endogamy and consanguinity will be required to fully evaluate this apparent property of the $wLOD$ method.

Accuracy of the *wLOD* estimator

To evaluate the sensitivity and specificity of the $wLOD$ method to detect ROA in dense genotype data, we simulated 50 independent replicates of genetic data under two demographic scenarios that are broadly representative of situations in which inbreeding and its effect on fitness are of interest as previously described [176] except that we considered a non-uniform distribution of recombination rates across the simulated chromosomes and allowed all base pairs to be mutatable (see **Methods**). Scenario 1 considered a small partially isolated population of constant effective size (N_e=75) that receives approximately one migrant per generation, simulated for 150 generations (4,350 years for a generation time of 29 years

[177]). Scenario 2 considered a medium sized closed population (N_e=500 simulated for 100 generations [2,900 years]). Each simulated dataset consisted of a single 250 Mb chromosome upon which ~750,000 polymorphic single-nucleotide variants (SNVs) segregate, consistent with the SNV density and length of chromosome 1 in The 1000 Genomes Project Phase 3 WGS data. The simulated WGS datasets used in downstream analyses contained 50 randomly chosen individuals from the final generation with genotypes for 709,862-746,963 SNVs in scenario 1 and 737,957-748,572 SNVs in scenario 2.

To better mimic real genetic datasets, we randomly introduced genotyping errors separately into each simulated dataset at a rate of 0.001, a conservative value that is similar to but slightly higher than the average rate of genotype discordance across 1,693 individuals between genotypes in their WGS data and those obtained at the exact same SNVs with the Illumina HumanOmni2.5 BeadChip [155], and we set ε to this value in all analyses. Analysis of the simulated pedigrees found the parents of individuals in the final generation to have a common ancestor on average three generations in the past for scenario 1 (all between 1 to 5 generations) and four generations in the past for scenario 2 (all between 1 to 7 generations) and M was set to these average values when analyzing their respective datasets.

Separately for each simulated dataset, we applied the $wLOD$ estimator considering windows of between 50 and 500 SNPs (in 10 SNP increments), count estimates of allele frequencies calculated using all 75 individuals, and the genetic and physical map positions of each genotyped position returned by the simulation program. All windows with $wLOD$ scores higher than the location of the minimum between the non-autozygous and autozygous modes in the $wLOD$ score distribution were considered autozygous [18]; overlapping autozygous windows were joined to define ROA. Here, we varied the proportion of overlapping windows that must be called as autozygous when defining ROA between 0 and 50 percent (in 1% increments). As each SNV is included in multiple windows (i.e. an SNV is included in 50 different windows at a window size of 50), near the edges of a true ROA some SNV will be included in both autozygous and non-autozygous windows as the sliding window enters and leaves the ROA. Requiring an SNV to be covered by a certain proportion of autozygous windows before it is placed within an ROA can improve the accuracy of ROA inferences when using a sliding-window approach [146].

For each simulated dataset, we then calculated three measures of how well inferred ROA agreed with true ROA reported by the simulation program. First, we calculated the power of the $wLOD$ method to detect true ROA, defined here as the total length of true ROA that is overlapped by inferred ROA divided by the total length of true ROA. Second, we calculated its false positive rate as the total length of inferred ROA that does not overlap with true ROA divided by the total length of inferred ROA. Finally, for all true ROA detected with the $wLOD$ method, we calculated the ratio of inferred ROA length and true ROA length for all ROA. Here, ratios greater than one indicates a tendency to overcall ROA by falsely including non-autozygous regions near the boundaries of the true ROA, while ratios below one indicate a tendency to instead undercall an ROA by falsely excluding true autozygous regions near the boundaries of the true ROA [178].

As can be seen in **Figure 4A**, large numbers of false positive ROA calls are made by the $wLOD$ method with a window size of 50 SNPs, decreasing markedly as the window size and the proportion of overlapping windows required during ROA construction increases. These patterns are consistent with the observation that false positive ROA calls are very small**—**on average 16.97 kb (standard deviation [SD] = 3.85) with a window size of 50 SNPs**—**and therefore delineated by a few erroneous autozygous windows that progressively fail to meet the required threshold during ROA calling as the window overlap fraction increases. Once window size reaches ~90 SNPs, the $wLOD$ estimator is able to distinguish autozygosity from homozygosity-by-chance with great precision. Conversely, numbers of false negative ROA calls increase as a function of window size and overlap fraction (**Figure 4B**). These patterns are consistent with the expectation that as window size increases smaller ROA increasingly go undetected (Additional File 1: **Figure S5A**), likely as a result of them being spanned by progressively fewer but larger windows and their autozygosity signal being increasingly masked by the inclusion of non-autozygous flanking regions in the $wLOD$ score calculation. Similarly, higher overlap fractions also lead to small ROA spanned by just a small number of autozygous windows increasingly going undetected (Additional File 1: **Figure S5D**) as they fail to meet the required threshold. Nevertheless, overall power to detect ROA with the $wLOD$

method only decreases slightly as window size and overlap fraction increase (**Figure 4C**), consistent with the expectation that at larger window sizes (Additional File 1: **Figure S5B**) and overlap fractions (Additional File 1: **Figure S5E**) the sliding window approach will have increasing difficulty in detecting smaller ROA but nonetheless retains high power to detect longer ROA. Finally, ratios of inferred to true ROA length increase as a function of window size and decrease as a function of overlap fraction (**Figure 4D**), reflecting the tendency of the $wLOD$ method to overcall the boundaries of smaller ROA at larger window sizes (Additional File 1: **Figure S5C**) and smaller overlap fractions (Additional File 1: **Figure S5F**) with those of longer ROA affected to a much lesser extent. All together, these patterns suggest that a suitable point within the parameter space at which to evaluate the sensitivity and specificity of the $wLOD$ method will be the smallest window size and overlap fraction combination at which no false-positive ROA are inferred and the average ratio of inferred to true ROA length is approximately one (**Table 3**), striking a balance between sensitivity to detect smaller ROA and the overall accuracy of ROA calls.

To evaluate how SNV density influences the sensitivity and accuracy of ROA inference with the $wLOD$ method we created three subsets of the simulated WGS datasets that reflect the SNV densities of commonly used human microarray-based genotyping platforms: Illumina's HumanOmni2.5-8 (125,000 SNVs) and OmniExpress-24 (50,000 SNVs) BeadChips and Affymetrix's Genome-Wide Human SNP 6.0 Microarray (80,000 SNVs). In addition, we included subsets with SNV densities consistent with the genotyping platforms used by ROA studies in cattle and dogs: Illumina's Bovine HD (80,000 SNVs) and Canine HD (18,000 SNVs) BeadChips. After the removal of monomorphic SNVs, the 125K, 80K, 50K, and 18K subsets contained between 117113-123766, 74953-79211, 46846-49507, and 16865-17823 polymorphic SNVs, respectively, for scenario 1, and between 121833-122815, 77973-78602, 48733-49126, and 17544-17686 polymorphic SNVs for scenario 2. ROA were inferred and evaluated exactly as described above for the WGS datasets containing ~750K SNVs, with the optimal window size and overlap fraction determined separately for each SNV density and demographic scenario (**Table 3**). Interestingly, optimal window size varied only slightly across the different SNV densities, lying between 60–130 SNPs and 70–120 SNPs for scenarios 1 and 2, respectively, but nevertheless increasing as a function of SNV density. The optimal window overlap fraction did however vary more widely, increasing as a function of SNV density and lying between 7–37% and 5–32% for scenarios 1 and 2, respectively.

As would be expected, the power of the $wLOD$ method to detect ROA increases as a function of ROA length and the density of SNV in the genetic dataset (**Figure 5**). While ROA longer than 1 Mb are captured extremely well (>99.7%) at all SNV densities explored, the detection of ROA shorter than ~1 Mb decreases appreciably as a function of SNV density. Nevertheless, even with only ~18,000 SNVs (1 SNV every ~14 kb) the $wLOD$ method is able to capture 96.3% and 89.0% of ROA under scenarios 1 and 2, respectively, with this increasing to 99.9% for both scenarios with 750,000 SNVs (1 SNV every ~333 bp). However, false discovery rates do increase dramatically with decreasing SNV density, particularly for smaller ROA (**Figure 5**) where they jump from 0.0045 and 0.0069 with 750,000 SNVs to 0.0445 and 0.1362 with 18,000 SNVs for scenarios 1 and 2, respectively, while longer ROA are much less affected: 0.0010 and 0.0001 with 750,000 SNVs increasing to 0.0200 and 0.0495 with 18,000 SNVs for ROA ≥ 5 Mb, respectively. It should be noted that these false discovery rates are solely the result of overcalling true ROA and not erroneous ROA calls. This is reflected in the ratios of inferred to true ROA length (**Figure 5**) that increase with decreasing SNV density, particularly for smaller ROA, and approach**—**but never quite reach**—**one with increasing ROA length.

Overall, these findings indicate that the $wLOD$ method is well powered to detect ROA with high sensitivity and good specificity at a wide range of SNV densities that are consistent with WGS as well as popular microarray-based platforms that are commonly used in human and non-human studies of ROA, and in particular long ROA that are of interest in studies of Mendelian and complex diseases and traits. In the simulations, both the optimal window size and the optimal overlap fraction increased logarithmically as a function of SNV density (R^2=0.9814 and R^2=0.8868, respectively, when considering their averages across scenarios). Fitting these averages against the natural logarithm of average SNV density D across all 50 replicates of their respective SNV subset, this suggests that as a rule of thumb future studies apply the $wLOD$ method at a window size equal to $16.400 \times log_e(D) + 218.020$ and an overlap fraction equal to

$0.0736\times log_e(D) + 0.8063$. Based upon these equations, and calculating SNV density as the number of autosomal SNPs on the microarray divided by the total length of the target species' autosomal genome, guideline settings for window size and overlap fraction with the commonly used human and non-human genotyping microarrays are: 111 SNPs (33%), 103 SNPs (29%), 85 SNP (21%), 81 SNPs (19%), and 59 SNP (9%) for Illumina's HumanOmni5, HumanOmni2.5, Bovine HD, OmniExpress, and Canine HD BeadChips, respectively, and 85 SNPs (21%) for the Affymetrix Genome-Wide Human SNP 6.0 Microarray. Considering the range of autosomal SNVs observed in the WGS data available for the 26 worldwide populations in Phase 3 of The 1000 Genomes Project (12–24 million SNV [155]) a window size of 128–140 SNPs and an overlap fraction of 40–45% would be recommended for WGS datasets. Nevertheless, the modest effect window size has on power to detect longer ROA across the simulated SNV densities (Additional File 1: **Figure S6**) would suggest that the use of more conservative (i.e. larger) window sizes will not greatly impact the ability of future studies to detect longer ROA of interest regardless of the source and density of the SNV data being analyzed. The window overlap fraction used in ROA construction can then be tailored to meet the needs to detect shorter ROA (Additional File 1: **Figure S7**) and to accurately place ROA boundaries (Additional File 1: **Figure S8**), where less restrictive (i.e. smaller) fractions can greatly improve the detection of shorter ROA without significantly impacting the accuracy of longer ROA inferences.

Performance of $wLOD$ against existing ROA detection methods

We have shown the $wLOD$ method to be well powered to detect ROA in genetic datasets consistent with WGS and microarray-based genotyping. We next evaluated how the power and false discovery rate of the $wLOD$ method compared with those of the original LOD method as well as the naïve genotype counting method implemented in *PLINK* [146] and the recently reported hidden Markov model (HMM) method implemented in the *RoH* function of *BCFtools* [154] using the datasets simulated above. We do not consider here the ROA detection methods of *GERMLINE* [148] and *Beagle* [179] as they have been previously shown to underperform compared with the method implemented in *PLINK* [149]. Since the false discovery and boundary placement properties of the sliding-window-based LOD and *PLINK* methods would be expected to differ from those of the $wLOD$ method due to their different underlying models, separately for each dataset we identified the optimal window size and overlap fraction for the LOD method (Additional File 2: **Table S1**) and *PLINK* (always a window size of 50 SNPs and an overlap fraction of zero) as described above. For *PLINK* we allowed at most 2% of SNPs to have heterozygous genotypes and 5% of SNPs to have missing genotypes for a window to be inferred to be autozygous [149]. The LOD method and *BCFtools/RoH* were applied using the same allele frequency estimates and error rate ε as the $wLOD$ method, while *BCFtools/RoH* additionally incorporated genetic map positions and performed Viterbi training with initial transition probabilities between autozygous and non-autozygous states and vice versa of 6.6×10^{-8} and 5.0×10^{-9}, respectively, to optimize its underlying model prior to ROA calling [154].

For both scenario 1 and 2, all four methods were able to detect >99.5% of ROA on average with 750,000 SNVs (**Figure 6A** and **6D**, respectively), representative of the density of SNVs observed in WGS data. Nevertheless, the $wLOD$ method outperformed both the original LOD method as well as *PLINK* and *BCFtools/RoH*, particularly at shorter ROA lengths. Interestingly, power with *BCFtools/RoH* became increasingly erratic at longer ROA lengths, most noticeably in scenario 1 (small isolated populations), for reasons that remain enigmatic. However, while the $wLOD$ method had a lower false discovery rate than the LOD method, it was notably higher than that of *BCFtools/RoH* and *PLINK*. Again, it should be noted that this elevated false positive rate solely reflects the overcalling true ROA due to the sliding-window approach employed and not erroneous ROA calls, with such overcalling easily reduced through the use of a more stringent overlap fraction but at the expense of power to detect short ROA. Nevertheless, average ratios of inferred to true ROA length were broadly similar across the wLOD, LOD, and *BCFtools/RoH* methods, where they are highest for extremely short ROA and decrease exponentially with increasing ROA length until they approach—but never quite reach—one, although ratios with *BCFtools/RoH* were marginally lower than those with the $wLOD$ and LOD methods in scenario 2. Conversely, average ratios with *PLINK* decreased noticeably as a function of ROA length**—**reaching 0.47 in scenario 1 and 0.81 in scenario 2**—**consistent with the expectation that as a consequence of its naïve model, *PLINK* will have a tendency to undercall ROA or return fragmented ROH calls across their span as a function of the distribution of heterozygous genotypes within the ROA, which would be expected to be most numerous near its boundaries. Overall, these observations would suggest that model improvements implemented in the $wLOD$ estimator (equation 2) that account for the confounding effects of LD, recombination, and mutation in the autozygosity likelihood calculation provide improved sensitivity and specificity in ROA calling over the original LOD estimator (equation 1). Additionally, they indicate that the $wLOD$ method's sliding window approach, which combines evidence for autozygosity across multiple SNVs, provides improved sensitivity to detect ROA compared with the HMM method of *BCFtools/RoH*, albeit with slightly decreased accuracy in ROA boundary placement.

When we consider simulated datasets consistent with those of genotyping microarrays we observe similar patterns to those observed with 750,000 SNVs (**Figures 6** and **S9** [Additional File 1]). For both scenarios 1 and 2, the $wLOD$ method consistently outperforms the LOD method as well as *BCFtools/RoH* and *PLINK* in terms of power, particularly at shorter ROA lengths. False discovery rates with the $wLOD$ method are consistently lower than those with the LOD method but remain slightly higher than those with *BCFtools/RoH*, while ratios of inferred to true ROA length remain similar across the $wLOD$ and LOD methods and *BCFtools/RoH*. As SNV density decreases from 750,000 SNVs down to 18,000 SNVs several patterns emerge. First, the difference in power between the $wLOD$ and LOD methods decreases as a function of SNV density (Additional File 1: **Figure S9B** and **S9D**), disappearing faster under scenario 2 (large closed populations) than under scenario 1 (small partially isolated populations). These patterns are consistent with the view that in datasets containing fewer SNVs, LD confounds the inference of ROA appreciably less than in datasets containing many SNVs. Consequently, the LD correction implemented in the $wLOD$ estimator (equation 2) increasing becomes less important as SNV density decreases, leading the LOD and $wLOD$ estimators to provide broadly similar autozygosity likelihoods. Nevertheless, false discovery rates with the $wLOD$ method are consistently lower than those with the LOD method, in agreement with the expectation that as SNV density decreases the probabilities of unobserved recombination and mutation events between genotyped SNVs increases, with the recombination and mutation corrections implemented in the $wLOD$ estimator (equation 2) enabling it to better account for these events than the LOD estimator (equation 1). Second, ratios of inferred to true ROA length with the *PLINK* method become more similar to those of the other three methods with decreasing SNV density. This pattern is consistent with the expectation that as SNV density decreases, the number of heterozygous genotypes within ROH will also decrease, allowing *PLINK* to increasingly detect the entire ROA. Finally, the performance of *BCFtools/RoH* decreases as a function of SNV density, although an appreciable loss of power only manifests when we reach 18,000 SNVs and is more pronounced in scenario 2 than in scenario 1 (Additional File 1: **Figure S9B** and **S9D**), suggesting that its HMM is sensitive to the effects of extended LD among sparsely distributed SNVs, a situation frequently encountered in closed populations due to elevated levels of general inbreeding. It should be noted, however, that *BCFtools/RoH* was designed for next-generation whole-genome and -exome data analysis and not for sparser microarray-derived genotype datasets, so its decline in performance in such datasets is to be somewhat expected.

Contrary to expectations based on frequent discrepancies in the autozygosity status of windows with the $wLOD$ and LOD estimators in The 1000 Genomes Project Phase 3 populations (**Figure 2**), in our simulated datasets the $wLOD$ method only provided modest improvements in power and false discovery rate over the original LOD method (**Figures 6** and **S9**). How can we reconcile the high similarity of ROA calls with the LOD and $wLOD$ methods in the simulated datasets with the appreciable differences in per-window autozygosity inferences made by their underlying estimators in The 1000 Genomes Project Phase 3 data? Considering the simulated datasets containing ~125,000 SNVs, which have a comparable SNV density to that of The 1000 Genomes Project Phase 3 Omni2.5 dataset investigated in **Figure 2**, and the same window size of 150 SNPs, across the 50 replicates for scenario 1 0.519% (SD=0.496) of windows were autozygous with the LOD estimator but not the $wLOD$ estimator, while 2.808% (SD=1.260) were autozygous with the $wLOD$ estimator but not the LOD estimator; for scenario 2 the values were 0.153% (SD=0.169) and 5.364% (SD=1.594), respectively. While the proportion of windows autozygous with the $wLOD$ estimator but not the LOD estimator in the simulated datasets is similar to that observed in The 1000 Genomes Project Phase 3 populations (**Figure 2D**), the proportion of windows autozygous with the LOD estimator but not the $wLOD$ estimator is about two orders of magnitude lower than the values observed in The 1000 Genomes Project Phase 3 populations (**Figures 2C**). Thus, while we observe the expected gain in sensitivity through a reduction in the contribution of occasional heterozygotes within ROH with the $wLOD$ estimator that enables improved detection of shorter ROA comprised of common haplotypes, we do not observe the expected inflation in LOD scores due to the confounding effects of LD among genotyped positions that leads to increased false positive ROA calls.

Based on their underlying models, we would expect the LOD (equation 1) and $wLOD$ (equation 2) estimators to provide highly similar inferences in situations where autozygosity patterns align almost perfectly with LD patterns among genotyped SNVs and are investigated with a sufficiently high density of SNVs that the probabilities of unobserved mutation and recombination events are effectively zero. The most parsimonious explanation for the surprisingly high similarity of ROA calls made by the LOD and $wLOD$ methods in the simulated datasets is therefore that LD patterns in these simulated datasets do not faithfully recapitulate the complexity of those found in real populations who have experienced much more complex histories than those simulated here, limiting the impact of the LD correction (equation 3) incorporated into the $wLOD$ estimator. We therefore expect to observe appreciably greater improvements in the sensitivity and specificity of ROA calls with the $wLOD$ method compared with the LOD method in real genetic data than in our simulated datasets.

Effect of genotyped SNV density on ROA detection in real data

We have shown the $wLOD$ method to be well powered to detect ROA in genetic datasets consistent with WGS and microarray-based genotyping, and to outperform a number of existing methods in terms of power although overcalling of ROA due to the sliding window approach it employs creates slightly higher rates of false discovery than a recently reported HMM model approach. While our simulations suggest that the $wLOD$ method has >99.8% power to detect ROA longer than 1 Mb across SNV densities that are consistent with those frequently used in human population- and disease-genetic studies (**Figure 6**), they do not capture the diversity of historical events and sociogenetic processes that have influenced genomic autozygosity patterns in contemporary worldwide human populations. Thus, we next sought to evaluate how robust ROA inferences are among genotype datasets created via WGS and whole-exome-sequencing (WES) as well as with the popular Illumina HumanOmni2.5-8 and OmniExpress-24 BeadChips using The 1000 Genomes Project Phase 3 data.

We first developed a WGS dataset comprised of all 75,071,695 SNVs that passed our quality control criteria (see **Methods**). Next, we developed a subset of the WGS dataset that was restricted to only the 1,830,512 SNVs that are located within the genomic regions captured by the Roche Nimblegen SeqCap EZ Human Exome Library v3.0 system to mimic a whole-exome-sequencing (WES) dataset ("WES dataset" henceforth). Finally, we developed a subset of the Omni2.5 dataset that was comprised of the 676,445 SNPs that are also present on the Illumina OmniExpress-24 BeadChip ("OmniExpress dataset" henceforth). As the $wLOD$ method explicitly accounts for LD among genotyped positions within a given window (equation 3) we do not consider LD pruned datasets. Similarly, since homozygosity for minor alleles at low to rare frequencies in the population is most informative for autozygosity inference with the $wLOD$ estimator (Additional File 1: **Figure S1A**), we also do not consider a minor allele frequency (MAF) pruned datasets.

For the WGS, Omni2.5 and OmniExpress datasets we applied the $wLOD$ method at the window size and overlap fraction suggested by our simulation analyses given their average SNV density across populations: 125 SNPs (40%), 95 SNPs (25%), and 80 SNPs (18%), respectively. As the SNV density of the WES dataset closely resembles that of the WGS dataset in the genomic regions it covers, we used the same window size and overlap fraction settings in both the WES and WGS datasets. For all datasets, μ was set to 1.18×10^{-8} [160] and M was set to seven, a conservative value broadly reflecting the average of effective population size estimates for populations included in The 1000 Genome Project [155,158,161]. For the Omni2.5 and OmniExpress datasets ε was set to 4.71×10^{-4}, the average rate of discordance across samples between genotypes in our Omni2.5 dataset and those obtained for 1,693 of the 2,436 individuals directly with the Illumina HumanOmni2.5 BeadChip [155], while in the WGS and WES datasets ε was instead set separately for each genotype as one minus its reported likelihood. This has the potential to improve the accuracy of ROA calls in NGS datasets by incorporating the uncertainty of each genotype call into the $wLOD$ score calculation, an important potential source of erroneous ROA calls in the context of their often higher and more variable per-genotype error rates compared with microarray-derived datasets [151,152]. As such, autozygous windows comprised of SNVs with low quality genotypes have a greater chance of being false-positive signals than those with higher quality genotypes, while low quality heterozygous genotypes—that in one possibility may be genotype calling errors—located in runs of higher quality homozygous genotypes have the potential to mask true autozygous signals.

For each dataset and population, we defined a $wLOD$ score autozygosity threshold as the location of the minimum between the non-autozygous and autozygous modes in its $wLOD$ score distribution [18]. Sample size was not observed to appreciably influence the location of the minimum between the non-autozygous and autozygous modes (Additional File 1: **Figure S10**). However, across 100 random samples of individuals greater consistency in its determination was observed with increasing sample size, particularly compared with sample sizes of less than 10 individuals, indicating that 10 or more individuals should be used to ensure a robust estimate of the threshold is obtained. All windows with $wLOD$ scores above threshold were considered autozygous [18], and overlapping autozygous windows were joined to define ROA contingent on the window overlap fraction used for that dataset.

Comparing ROA identified in the WGS and Omni2.5 datasets, we find Omni2.5 ROA to be frequently longer than their corresponding WGS ROA and in most cases to completely encompass the WGS ROA (**Figure 7A**). The magnitude of their length discrepancies decreases with increasing ROA length, consistent with the expected effects of decreased SNV density on the accuracy of inferred ROA boundaries. In addition, while all Omni2.5 ROA are present in the set of WGS ROA, the reverse is not true (**Figure 7B**). Many short ROA (<500 kb) detected in the WGS dataset are not found in the Omni2.5 dataset, with the fraction of missing ROA decreasing with increasing distance from Africa, reflecting the effect of increasing LD [162,163] on our ability to detect shorter ROA with the sparser set of SNVs in the Omni2.5 dataset. Concordance between the WGS and Omni2.5 datasets for intermediate (500 kb to 1.5 Mb) and long (>1.5 Mb) ROA is generally high, although in many populations the fraction of WGS ROA missing in the set of Omni2.5 ROA remains nontrivial. These fractions generally increase as a function of distance from Africa, likely reflecting the reduction in haplotype diversity with decreasing genetic diversity [95,164-167] decreasing our ability to distinguish autozygosity from homozygosity-by-chance, particularly over extended genomic regions when genotypes are only available for a fixed set of SNVs that were selected for their generally high level of polymorphism worldwide.

Similar patterns are observed when we compare ROA identified in the Omni2.5 and OmniExp datasets, where almost all OmniExp ROA are present in the set of Omni2.5 ROA (Additional File 1: **Figure S11B**) and encompass their generally shorter corresponding Omni2.5 ROA (Additional File 1: **Figure S11A**). While many short ROA detected in the Omni2.5 dataset are not found in the OmniExp dataset, both intermediate and long ROA are captured extremely consistently between the two datasets despite their different SNV densities. Likewise, when we compare ROA identified in the WGS and WES datasets, almost all WES ROA are present in the set of WGS ROA (Additional File 1: **Figure S12B**) and tend to encompass their generally shorter corresponding WGS ROA (Additional File 1: **Figure S12A**). However, while numbers of short and intermediate ROA identified in the WGS dataset but not the WES dataset are much higher than in the same comparison between the WGS and Omni2.5 datasets (**Figure 7**), the numbers of long ROA identified in the WGS dataset but not the WES dataset are instead similar. This indicates that the non-uniform and often sparse distribution of SNVs in the WES dataset does not impact the detection of long ROA more than would be expected following a general reduction in SNV density.

Overall, these findings are consistent with the higher density of SNVs in the WGS dataset and the presence of many more rare and low-frequency SNVs detected by NGS compared with microarray-based genotyping platforms—which are particularly informative about autozygosity under our likelihood model (Additional File 1: **Figure S1A**)—greatly improving our ability to detect ROA. Nevertheless, many long ROA that are of interest in Mendelian and complex disease studies are well captured by the sets of SNVs included on Illumina's HumanOmni2.5-8 and OmniExpress-24 BeadChips. However, the sparse and non-uniform genomic distribution of SNVs in the WES dataset creates difficulties when inferring short and intermediate ROA with the $wLOD$ method, despite the presence of rare and low-frequency SNVs, while long ROA are instead captured almost as well as with genotyping microarrays. We therefore do not recommend using the $wLOD$ method to detect ROA in WES datasets generated by future studies.

Classification of ROA

ROA of different lengths reflect homozygosity for haplotypes inherited IBD from common ancestors at different depths in an individual's genealogy: longer ROA most likely arise due to recent ancestors and shorter ROA due to more distant ancestors. We previously advocated that ROA be classified into *G* length-based classes using the Gaussian mixture model approach applied on their physical map lengths (in bp) that groups ROA based upon their supposed ages [18]: (A) short ROA that measure tens of kilobases and that are of the length at which baseline patterns in LD in a population produce autozygosity through the pairing of two copies of the same ancient haplotype, (B) intermediate length ROA that measure hundreds of kilobases to several Mb and that are likely the result of background relatedness—recent but unknown kinship between parents due to limited effective population sizes—and (C) long ROA that measure multiple megabases and are likely the result of recent parental relatedness. The choice of $G = 3$ was motivated by the observation that at $G > 3$, the additional classes were not discrete; that is, they were encompassed by one of the existing classes (Additional File 1: **Figures S13A** and **S13C**).

This classification approach is limited by the imperfect correlation between physical map lengths and genetic map lengths (Additional File 1: **Figure S14**), a more accurate representation of the relationship between ROA length and age [180,181] that is not biased by the non-uniform genomic distribution of recombination rates [182]. If we instead classify ROA based on their genetic map length (in cM) using a Gaussian mixture model we find that regardless of the number of classes considered they are always discrete (Additional File 1: **Figures S13B** and **S13D**). This would suggest that the original loss of discreteness when classifying based upon physical map length may reflect the confounding effects of physically long but genetically short (and vice versa) ROA on the overall length distribution. Nevertheless, regardless of whether physical or genetic map lengths are used the overall pattern of fit with increasing class number remains highly similar (Additional File 1: **Figures S13A** and **S13B**, respectively), where Bayesian Information Criterion (BIC) likelihoods plateau at around $G = 5$ with the WGS and Omni2.5 data and at around $G = 4$ classes with the OmniExpress data (not shown). The smaller class number for the OmniExpress dataset compared with the WGS and Omni2.5 datasets is consistent with the expectation that smaller ROA will be poorly captured by its sparser set of SNVs, ultimately leading to the loss of the shortest ROA class detected in the WGS and Omni2.5 datasets. Note that for all populations the maximum BIC likelihood is reached at $G > 5$. Future studies investigating fine scale ROA patterns may wish to consider values of *G* at which BIC is maximized, however for illustrative purposes we consider $G = 5$ here since the increase in BIC at $G > 5$ is small.

When considering a five class classification scheme, the longest class ($G = 5$) contains ROA that likely arise from recent parental relatedness and the penultimate longest class ($G = 4$) contains ROA that likely arise from recent population processes, while the shortest classes ($G = 1$-3) contain ROA arising through the pairing of two copies of much older haplotypes that have common ancestors at different times in the distant past. Sample size was observed to have a greater effect on ROA classification (Additional File 1: **Figure S15**) than on $wLOD$ score threshold (Additional File 1: **Figure S10**), with the proportion of ROA whose classification differed from that assigned when all available individuals are used decreasing as a function of sample size. Importantly, the proportion of misclassified ROA decreases with increasing ROA class, with those in the longest class ($G = 5$) infrequently misclassified (mean=0.052 with SD=0.029 across all 26 populations at a sample size of 25) while those in shorter classes were more frequently affected (mean=0.092 with SD=0.046, mean=0.091 with SD=0.045, mean=0.083 with SD=0.045, and mean=0.068 with SD=0.042, for G = 4 to 1, respectively). These observations indicate that sample size is an important factor when classifying ROA using a Gaussian mixture model, but in general samples sizes of at least 25 individuals should provide reasonably robust classification of ROA using this approach, particularly longer ROA that are of interest in genetic studies on Mendelian and complex diseases.

Geographic patterns in ROA

We have shown the $wLOD$ method to be well powered to detect ROA in genetic datasets consistent with WGS and microarray-based genotyping, while our investigation of a Gaussian mixture model approach for ROA classification based upon their genetic map lengths indicates the presence of five ROA classes in The 1000 Genomes Project Phase 3 populations, a higher number than was used in our earlier study of the Human Genome Diversity Panel (HGDP) and International HapMap Project (Phase 3) populations that used a microarray-derived dataset and classified ROA based upon their physical map lengths [18]. To evaluate how genome-wide patterns in ROA inferred with the $wLOD$ method and classified into five classes via a Gaussian mixture model applied to their genetic map lengths compared with those of earlier studies, we performed the first high-resolution survey of ROA patterns in The 1000 Genomes Project Phase 3 populations based upon ROA inferred in the WGS dataset as described above.

Consistent with previous studies [12,18,22], ROA of different lengths have different continental patterns among the 26 populations included in Phase 3 of The 1000 Genomes Project both with regards to their total lengths (**Figure 8**) in individual genomes as well as in their non-uniform distributions across the genome (**Figure 9**) that are correlated with spatially variable genomic properties such as recombination rate (Additional File 1: **Figure S16**) and signals of natural selection (Additional File 1: **Figure S17**), reflecting the distinct forces generating ROA of different lengths. Total lengths and numbers of ROA in the shortest (G = 1-3) and to some extent intermediate (G = 4) classes increase with distance from Africa, rising in a stepwise fashion in successive continental groups (**Figures 8**), in agreement with the observed reduction in haplotype diversity with increasing distance from Africa [162,183-185]. Those of the longest class (G = 5) do not show a similar stepwise pattern, instead exhibiting higher and more variable values in populations where consanguinity in more frequent (**Table 2**) and inbreeding coefficient estimates are generally higher [186]. Notably, the East Asian Dai have remarkably high total lengths of short ROA (G = 1-3), potentially reflecting their small population size**—**~1.2 million in Yunnan province, China [187], where The 1000 Genomes Project samples were collected**—**and complex evolutionary history [188,189].

Recombination and natural selection

The strength of the correlation between the genomic distribution of ROA and recombination rate decreases with increasing ROA class (Additional File 1: **Figure S16**), consistent with the expectation that the patterns of genetically shorter ROA will be determined by recombination to a greater extent than longer ROA, which due to their more recent origins have had fewer opportunities for recombination events to systematically influence their patterns. Conversely, the correlation between ROA patterns and signatures of natural selection is strongest for class 2-3 ROA, and to some extent intermediate class 4 ROA, while it is very weak for the shortest (G = 1) and longest (G = 5) ROA classes (Additional File 1: **Figure S17**). These patterns are compatible with natural selection having primarily influenced genomic diversity patterns in the distant past, with autozygosity for the relics of the haplotypes that arose during those events manifesting as class 1-4 ROA, dependent upon how long ago the event occurred.

The long term effects of natural selection on patterns of ROA might be expected to be most evident in genomic regions encompassing genes implicated in one or more Mendelian diseases, where purifying selection acting on strongly deleterious alleles, which may occur more frequently in such genes due to their apparent importance for human health, would be expected to increase levels of homozygosity relative to genes much less frequently subjected to purifying selection. Using the union of two previously reported lists of genes associated with autosomal dominant (669) and recessive (1130) diseases in the Online Mendelian Inheritance of Man (OMIM) database [190-192], we created a list containing genes not associated with autosomal dominant or recessive diseases (24,260; "non-OMIM" henceforth); genes associated with both autosomal dominant and recessive diseases were ignored. For each individual, we then calculated the fraction of the total lengths of all autosomal dominant, autosomal recessive, or non-OMIM transcribed regions that are overlapped by ROA based on their genomic positions in build HG19 of the University of California – Santa Cruz (UCSC) reference genome assembly. Strikingly, regardless of the ROA length class considered, the fraction for OMIM dominant genes was almost always higher than that of non-OMIM genes (P<10^{-16} in all comparisons; Wilcoxon signed rank test), while the opposite was true for OMIM recessive genes (P<10^{-16} in all comparisons; Additional File 1: **Figure S18**). Nevertheless, the pattern is strongest for intermediate length ROA classes (G = 2−4) and weakest for the shortest (G = 1) and longest (G = 5) classes. Together, these results are compatible with deleterious alleles occurring less frequently in non-OMIM genes than in OMIM dominant genes, where they are efficiently removed from the population via purifying selection acting on both their homozygous and heterozygous forms, creating increased autozygosity at lengths consistent with population-level processes rather than inbreeding. One possible explanation for the decreased autozygosity around OMIM recessive genes compared with non-OMIM genes would be increased embryonic lethality and/or childhood mortality with individuals homozygous for deleterious recessive mutations in OMIM recessive genes, leading to reduced autozygosity in genomic regions encompassing them in the extant population.

Genes that have been the target of positive selection might be expected to reside within genomic regions that are more frequently autozygous in the general population than those harboring genes that have not. Considering the fraction of each gene's transcribed region that is in a ROA in each individual's genome, we compared their median fraction across individuals in each population (Additional File 1: **Figure S19**). While most genes have a median fraction of about zero, a number of genes that lie within genomic regions spanned by ROA in more than 90% of individuals in a population. Across populations, we observe 54 such instances with long class 5 ROA that represent seven distinct genomic regions (Additional File 2: **Table S2**), 159 with intermediate length class 4 ROA (22 distinct regions; Additional File 2: **Table S3**), and 31 (nine distinct regions; Additional File 2: **Table S4**), seven (five distinct regions; Additional File 2: **Table S5**), and 480 (46 distinct regions; Additional File 2: **Table S6**) with short class 1–3 ROA, respectively. While most genes in these regions fall within the non-OMIM group, two of the genes enriched for class 4 ROA (*CFC1* and *SMN1*) and nine of the genes enriched for class 1 ROA (*SLC25A20*, *NDUFAF3*, *LAMB2*, *GPX1*, *NPRL2*, *ACY1*, *MRPS16*, *LCAT*, and *COX4I2*) are from the OMIM recessive group, while one gene enriched for class 1 ROA is from the OMIM dominant group (*THAP1*). Future investigation of genes that are unusually frequently overlapped by ROA in the general population may provide new insights into the role of recessive variation in human phenotypic diversity and common disease risk as well as the genes within which such variation acts.

Genomic distribution

Genomic distributions of shorter ROA (G = 1-4) are similar among populations from the same geographic region (Additional File 1: **Figures S20B-E**) and closely mirror the patterns of pairwise F_{ST} among populations (Additional File 1: **Figure S20A**; Procrustes similarity statistic t_0>0.803), while those of the longest ROA class (G = 5) vary more widely among populations (Additional File 1: **Figure S20F**; t_0=0.466). Overall, these patterns are consistent with the interpretation that shorter ROA (G = 1-4), for which neighboring populations have similar patterns, reflect autozygosity that arises through population processes on different evolutionary timescales, while longer ROA (G = 5), for which neighboring populations do not necessarily have similar patterns, reflect autozygosity that instead arises through more recent cultural processes such as inbreeding [18].

Autozygosity hotspots

The non-uniform genomic distribution of the different ROA classes and their variability among populations creates autozygosity hotspots that are in some instances shared among subsets of the populations. For example, there is a hotspot for class 4 ROA on the q-arm of chromosome 2 that is common to three of the five European populations and encompasses the human lactase gene (*LCT*; **Figure 10**) that was not detected in our original study of the HGDP and HapMap populations that included 10 from Europe [18]. In this genomic region, we observe high frequencies of intermediate length class 4 ROA in the Northern European FIN and GBR populations as well as the European American (CEU) group, but not in the Southern European TSI and IBS populations or any other population in the dataset. The presence and absence of this hotspot broadly reflects worldwide patterns in lactase persistence frequency [193,194]. Lactase persistence is most frequent in Northwestern Europe [195,196] where it is caused primarily by a single mutation in *LCT* that rose to high frequency as a consequence of natural selection in response to the rise of milk consumption and pastoralism [194,197,198]. It decreases in frequency through Eastern and Southern Europe and Central/South Asia reaching near-zero frequencies in East Asia and the Americas [193,196,199-201], while it is present to varying degrees in admixed Mestizo [202-204] and African American [199,204] populations as a consequence of their recent European ancestry. Thus, we observe high levels of autozygosity around *LCT* in the GBR, FIN, and CEU populations and markedly lower but noticeable levels in the IBS, but no observable signal in the TSI or any of the Asian or admixed populations. While lactase persistence is present at moderately high frequency in sub-Saharan Africa it is caused by several different mutations [194,205] and the African populations included in The 1000 Genomes Project are located predominantly in historically non-milking areas of the continent [197]. Consequently, we do not observe a similar autozygosity signal in the African populations as we do in the Northern European populations.

Interestingly, we also observe a hotspot for the longest ROA class (G=5) at the same location in the Northern European CEU and GBR populations ~770kb downstream of the *LCT* gene (**Figure 10**), while a weaker spike in class 5 ROA frequency is seen in the FIN population. This hotspot encompasses four genes within its core region (chr2:135,375,000-135,775,000) that encode a transmembrane protein (*TMEM163*), an aminocarboxymuconate semialdehyde decarboxylase (*ACMSD*), cyclin T2 (*CCNT2*), and a mitogen-activated protein kinase kinase kinase (*MAP3K19*). The maximum normalized haplotype-based selection statistic nS_L [206] score observed in the CEU, GBR, and FIN populations within the core region is 4.980, 4.818, and 4.962, respectively, suggesting that this ROA hotspot potentially reflects the outcome of recent positive selection. However, none of the genes within this hotspot are known to have functional consequences when mutated, leaving the cause of this ROA hotspot and its putative signals of positive selection enigmatic.

Overall, frequency patterns in this genomic region of the different ROA classes in the Northern European CEU, GBR, and FIN populations are consistent with positive selection having occurred at two different time-points. The extended haplotypes created by historical positive selection acting on the single *LCT* mutation that arose in ancestral Northern Europeans have, over subsequent generations, decreased appreciably in length, but due to the marked reduction in haplotype diversity in the surrounding region commonly create intermediate length class 4 ROA through background population processes. Conversely, the presence of extended IBD haplotypes creating longer class 5 ROA in a genomic region ~770 kb away from *LCT* would be compatible with positive selection acting much more recently, in agreement with the atypically high nS_L scores observed within this region in these populations.

Statistical inference of enrichment of autozygosity signals between groups

A unique feature of the *wLOD* ROA detection approach is the availability of log-likelihoods of autozygosity for each window in each individual examined. It is therefore possible to directly compare the strength of autozygosity signals between two or more groups of individuals to identify those windows that have significantly greater evidence for shared autozygosity signals in one group compared with the others [150]. In one possibility, such an approach could be used to identify genomic regions that have stronger signals of autozygosity in affected versus unaffected individuals and thus may harbor disease-associated mutations. Similarly, genomic regions with significantly stronger signals of autozygosity in one subset of a population compared to another other may reflect founder effects if there is limited gene flow between them or the presence of adaptive alleles in one subset but not the other that have risen to high frequency.

We demonstrate the principle of this approach using three of the five Central/South Asian groups included in Phase 3 of The 1000 Genomes Project who represent subpopulations within the larger Indian population: BEB, GIH, ITU, PJL, and STU. Genetic diversity patterns in these five groups support the presence of two genetically distinguishable clusters within the GIH, ITU, and PJL (Additional File 1:

Figure S21). When instead compared pairwise, the larger of the two ITU clusters lies intermediate between the smaller ITU cluster and the larger of the two GIH, PJL, or STU clusters, while the largest of the PJL clusters overlaps significantly with the smaller GIH cluster (not shown). The GIH individuals were sampled in Houston, TX, while the BEB, ITU, PJL, and STU individuals were all sampled in the UK. Given the intermediate locations of the larger ITU and PJL clusters in the pairwise comparisons, they may potentially reflect admixed individuals within these sample sets. However, both clusters are tightly bunched arguing against this possibility given the normal dispersion of admixed individuals in such analyses owing to their continuum of admixture levels [207,208]. In another possibility, these distinct clusters might represent the unintentional sampling of distinct endogamic communities whose restrictive marital practices under the long-established Indian caste system has made them distinguishable genetically [209].

Because we would expect differential autozygosity signals among groups to have arisen relatively recently through population or cultural processes, window size is not constrained by our power to detect shorter, more ancient, ROA. A natural window size to use when searching for differential autozygosity signals between groups is therefore the one whose $wLOD$ score distribution can best discriminate between autozygous and non-autozygous windows. In one possibility, this can be defined as the window size that maximizes the distance between the autozygous and non-autozygous modes—measured here as the distance between the modal score in each mode (**Figures 1B** and **S2** [Additional File 1]). Using the WGS dataset and optimal window sizes of 450, 580, and 610 SNVs for the GIH, PJL, and ITU, respectively, we compared the $wLOD$ scores of individuals present in each of their two clusters (Additional File 1: **Figure S21**) and evaluated the significance of their observed differences with the permutation-based approach described in Wang *et al*. [150] except that here we use a Wilcoxon rank-sum test instead of the two sample *t*-test suggested by Wang *et al*. as it is much less sensitive to the presence of outliers but has similar power to detect a location shift [210]. Briefly, separately for each group, we first create a distribution of test statistics under the null hypothesis of no difference in $wLOD$ scores between clusters using 1,000 permutations of cluster labels, recording for each permutation the maximum observed test statistic across all windows genome-wide. Next, separately for each window, a genome-wide adjusted *P*-value for the significance of the observed differences in $wLOD$ scores between clusters is then calculated as the proportion of the maximum genome-wide test statistics observed in the 1,000 permutations that exceeded the test statistic obtained with the true labels for that window. Finally, for each cluster, genomic regions enriched for autozygosity signals in that cluster compared with the other were defined by joining together overlapping windows with a permutation *P*-value (P_{perm}) ≤ 0.05.

Intriguingly, while we would not *a priori* expect to observe significant differences in the strength of autozygosity signals between the two apparent clusters within the GIH, ITU, and PJL sample sets, we did identify one genomic region significantly enriched for autozygosity signals in cluster A compared with cluster B in both the ITU and PJL (**Figures 11B** and **11C**; **Table 4**); no regions were identified in the GIH (**Figure 11A**). The genomic region in the ITU lies within the transcription elongation regulator 1 like (*TCERG1L*) gene that has been associated with regulation of plasma levels of the adipokine adiponectin [211], a modulator of glucose regulation and fatty acid oxidation [212] implicated in obesity, diabetes, coronary artery disease and Crohn’s disease risk [213-215]. The genomic region in the PJL encompasses the transmembrane phosphoinositide 3-phosphatase and tensin homolog 2 (*TPTE2*) gene, a paralog of the phosphatase and tensin homolog (*PTEN*) tumor suppressor [216] implicated in hepatic carcinogenesis [217] that has been found to harbor SNPs with significant allele frequency differences between males and females in European and African populations [218]. While the underlying basis for these differential autozygosity signals remains enigmatic in the absence of more detailed information on these individuals, their identification highlights the potential of our approach to identify genomic regions with differential autozygosity signals between groups that may reflect the presence of variants that have experienced differential selection histories or that influence differences in their predisposition to disease. Moreover, these findings highlight the need for further investigations among well-defined endogamic groups from India to facilitate our understanding of the genomic consequences of the long-established caste system.

Discussion

We have reported an improved likelihood-based estimator for the detection of ROA in genome-wide SNV genotype data derived from either microarray platforms or WGS that accounts for autocorrelation among genotyped positions and variability in the confidence of individual genotype calls as well as the probabilities of unobserved mutation and recombination events. Fully accounting for LD among SNVs in a given window is important, because in genomic regions of high LD many pairs of individuals will share common haplotypes that are homozygous identical-by-state but not ROA in the sense defined here (i.e., inherited IBD from a common ancestor). Thus, including such spurious windows would add noise when looking for ROA for the purpose of autozygosity mapping. The incorporation of LD in our model reduces false-positive ROA detection, affording us the ability to identify smaller ROA segments with greater fidelity. An alternative approach to accounting for LD is to prune the dataset prior to its analysis. However, such an approach first requires those SNV with MAF less than 5% to be removed, which would significantly reduce the power of the $wLOD$ method to detect ROA by removing those low-frequency and rare variants whose homozygosity is most indicative of autozygosity under its likelihood model (Additional File 1: **Figure S1A**). Further, such pruning cannot completely remove LD from the dataset being analyzed, with a pairwise r^2 threshold of 0.5 typically applied [149]. The incorporation of LD into the model therefore better controls for the autocorrelation of autozygosity signals among nearby SNV than is attainable with LD pruning, thereby improving the specificity of the ROA it detects particularly in regions of moderate to high LD.

Similarly, accounting for the probabilities of unobserved recombination and mutation events in the genomic interval spanned by the window becomes increasingly important as a function of inter-marker distance, particularly in situations where these probabilities become nontrivial such as in lower-density microarray-derived genotype datasets. By modeling these probabilities based on the assumed number of generations since the last common ancestor of the apparent autozygous haplotypes, which we have set here based on the reported effective sizes of the populations included in The 1000 Genomes Project [155,158,161], we minimize the number of false positive ROA that can be erroneously inferred when recombination and mutations events onto very similar haplotype backgrounds give the appearance of autozygosity when paired with a non-recombined haplotype. An alternative approach would be to set an arbitrary maximum inter-marker distance allowed when calling ROA; dividing into two any inferred ROA that spans an inter-marker interval greater than that maximum. However, this has the potential to erroneously break-up long ROA, potentially impacting downstream analyses that use ROA length one of their filtering criteria. By incorporating mutation and recombination weightings into the $wLOD$ model we therefore take a more informed and less-biased approach to this issue, thereby improving the detection of longer ROA particularly in datasets containing sparser sets of SNVs.

We have shown the $wLOD$ ROA detection method to be well-powered to infer ROA in genetic datasets consistent with those generated by WGS and microarray-based genotyping. We recommend using this method together with a model-based ROA classification approach [18] based on genetic map lengths to distinguish ROA arising from population-level LD patterns on different evolutionary timescales (classes G = 1-4) from those arising from more recent cultural processes such as inbreeding (class G = 5). Our findings suggest that our detection approach is robust for analyses of as few as 10 individuals. However, model-based classification requires at least 25 individuals to provide a robust classification solution. Moreover, to ensure allele frequency and LD estimates used with the $wLOD$ estimator are close to their true value in the population, at least 30 unrelated individuals should ideally be used in their estimation [219,220]. Intriguingly, our observation of trimodal $wLOD$ score distributions for a subset of the 26 populations analyzed here, all known to practise both endogamy and consanguinity to varying degrees, suggests that this method may be able to distinguish autozygosity arising from different cultural processes that act on different time scales. Future work within well-defined endogamic and non-endogamic groups that practice consanguinity, as well as within simulated datasets exploring the breadth of possible isolation and inbreeding parameters observed in human populations, will be required to clarify this apparent property of the $wLOD$ method and evaluate its potential human genetics applications.

Comparisons of the ROA inferred using the $wLOD$ method on different microarray-derived and NGS datasets created from The 1000 Genomes Project Phase 3 WGS data suggest that long and to some extent intermediate length ROA are captured consistently by WGS and microarray-derived datasets. However, detection of shorter ROA does vary substantially among the different datasets as a consequence of the decreasing resolution and sensitivity attainable as the genome-wide density of genotyped positions decreases. An observation reflected in the notable lack of consistency between ROA inferred in the WES dataset and those identified in the WGS dataset. Nevertheless, population-genetic analyses of genomic ROA patterns among the 26 populations included in The 1000 Genomes Project on the basis of WGS data are consistent with our previous findings in the 64 worldwide populations included in the HGDP [221,222] and International HapMap Project [223] on the basis of ~600,000 microarray-derived SNP genotypes [18]. These observations would therefore suggest that ROA studies using microarray-derived genotype data have similar power to detect genomic ROA patterns, and in particular those of longer ROA that are of interest to the disease genetic community due to their enrichment of deleterious variation carried in homozygous form [96,97], as those using WGS data.

We have compared the $wLOD$ method against a commonly used naïve genotype counting method implemented in the software *PLINK*, as well as the recently reported HMM method of the *BCFtools* software package, under two demographic scenarios in which ROA will be of interest in population- and disease-genetic studies. In our genetic simulations the *PLINK* approach performed surprisingly well, potentially reflecting their relatively short duration which limited the opportunities for new mutations to arise on the IBD haplotypes that ultimately underlied ROA in the final generation. Indeed, only ~4.01% and ~14.36% of SNVs in our simulated datasets were *de novo* mutations not present in the founder individuals under scenarios 1 and 2, respectively, while just ~2.14% and ~2.91% of SNVs had MAF < 5%. Conversely, across the 26 populations in The 1000 Genomes Project Phase 3 WGS data on average 56% of SNVs had MAF < 5%. Nevertheless, the $wLOD$ method had greater power to detect ROA versus *PLINK* across all SNV densities considered here. This difference reflects the very limited ability of the *PLINK* approach, which allows for only occasional missing or heterozygous genotypes when determining the status of a window to account for possible genotyping errors and mutations, to distinguish genomic regions that are homozygous-by-chance from those that are autozygous. In contrast, the $wLOD$ method incorporates population allele frequency and LD estimates and an assumed genotyping error rate as well as accounts for the probabilities of unobserved mutations and recombination events when inferring the autozygosity status of a window, enabling more rigorous assessments of the possibility of genotyping errors and the loss of information caused by missing data. In addition, it provides a more precise measure of the probability that a given window is truly autozygous rather than simply homozygous by chance. Thus, the greater power of the $wLOD$ method compared with *PLINK* reflects the greater number of false negative ROA expected under the naïve autozygosity model implemented in *PLINK*.

Comparisons of the $wLOD$ method with the recently reported *RoH* function of *BCFtools* have consistently shown it to have improved power to detect ROA, and smaller ROA in particular, across all SNV densities considered here, which are representative of WGS and microarray-based genotyping platforms. However, false discovery rates of the $wLOD$ method are slightly higher than those of *BCFtools/RoH*, wholly reflecting a more permissive placement of ROA boundaries marginally outside of their true locations as a consequence of the sliding window approach employed. While the underlying likelihood models of the $wLOD$ and *BCFtools/RoH* approaches are similar, there are two aspects of the $wLOD$ method that explain its higher power. First, by summing over all SNVs within a given window, the $wLOD$ method is better able to detect the autozygosity signals of ROA comprised of older (shorter) haplotypes whose constituent SNVs individually provide only weak to modest autozygosity support than the pointwise HMM employed by *BCFtools/RoH*. Second, the $wLOD$ method adjusts each SNV's log-likelihood by the probabilities that no unobserved recombination and mutation events have occurred in the interval between it and the preceding SNV in the last M generations (equation 2), where M is set based on the expected time since the most recent common ancestor in an individual's maternal and paternal lineages given the effective size of the population. *BCFtools/RoH* does not account for unobserved mutations in its inference model, and only allows for up to a single recombination event to have occurred within a given interval [154]. Thus, for longer ROA and those comprised of older haplotypes inherited IBD from an ancient ancestor, we would *a priori* expect *BCFtools/RoH* to have greater difficulty in making inferences as it will underestimate the number of recombination events that may have occurred as these haplotypes segregate in the general population. This may potentially underlie the noticeably erratic patterns observed with its power to detect ROA greater than 1.5 Mb in the higher SNV density simulated datasets (**Figure 6**).

Finally, the $wLOD$ method distinguishes itself from *BCFtools/RoH* and *PLINK* through its ability to directly detect genomic regions enriched for autozygosity signals in one population or group compared with one or more others without requiring the inference of ROA first. We have applied this approach within the Gujarati (GIH), Punjabi (PJL), and Telugu (ITU) Asian Indian groups, comparing $wLOD$ scores in two distinct clusters of individuals identified via multidimensional scaling of allele sharing dissimilarities (Additional File 1: **Figure S21**). We identified two genomic regions enriched for autozygosity signals in one of the two clusters, one in the ITU and another in the PJL, that contain genes implicated in the regulation of metabolism and the risk for developing liver cancer, respectively (**Table 4**). If we instead set a more permissive threshold of $P_{\text{perm}} \leq 0.1$ when defining enriched regions, we identify an additional seven genomic regions marginally enriched for autozygosity in one cluster compared with the other (Additional File 2: **Table S7**). One of the seven regions was identified on chromosome 2 in ITU cluster A and contains two genes: *G6PC2*, a pancreatic glucose-6-phosphatase implicated in the modulation of fasting plasma glucose levels [224] that is a major target of cell-mediated autoimmunity in diabetes [225], and the ATP-binding cassette transporter gene *ABCB11*, mutations in which cause autosomal recessive progressive familial intrahepatic cholestasis [226,227]. In addition, a region on chromosome 17 also identified in ITU cluster A contains seven genes that include *USH1G*, mutations in which cause autosomal recessive deafness in both humans [228,229] and mice [230,231]. Finally, a region on chromosome 16 identified in PJL cluster A contains four genes including the mechanically-activated ion channel gene *PIEZO1*, mutations in which cause autosomal recessive generalized lymphatic dysplasia [232,233] as well as autosomal dominant hemolytic anemia [234,235].

The presence of genes that cause autosomal recessive diseases in three of the seven marginally significant regions**—**a highly unlikely observation (P<0.008 across 1,000 random draws of genomic regions of equivalent size)**—**suggests the intriguing possibility that, if these clusters do indeed represent distinct endogamic communities, they may be the hallmark of cultural and selection processes related to the differential presence of deleterious genetic variants in these genes. Future comparative autozygosity analyses of well-defined endogamic communities within the different subpopulations of India considering much larger sample sizes than were available here will facilitate our understanding of the genomic consequences of the long-established caste system and further clarify its potential role in contributing to genetic predisposition in complex disease risk and negative health outcomes.

Conclusions

To facilitate community adoption of the $wLOD$ ROA detection method as well as classification based on genetic map length via a Gaussian mixture model, we have implemented these approaches in the software *GARLIC* (G*enomic* A*utozygosity* R*egions* L*ikelihood-based* I*nference and* C*lassification*) [178] that can be downloaded at https://github.com/szpiech/garlic. As a guide, analysis of the 97 individuals in the CEU population on a Dell Precision T7600 workstation running RedHat Enterprise Linux (v.7.3) with multi-threading support enabled (16 2.60 GHz threads total) took ~2½ minutes for the OmniExp dataset, ~6½ minutes for the Omni2.5 dataset, and ~40 minutes for the WGS dataset, and occupied at most ~3 Gb, ~7 Gb, and ~20 Gb of RAM, respectively. Future enhancements planned for *GARLIC*'s core engine are expected to significantly reduce its runtime and memory usage. We also provide a searchable online database of ROA identified in The 1000 Genomes Project Phase 3 populations as well as a ROA genome browser based on the *JBrowse* browser interface [236] in which to explore their genomic distribution with respect to various genomic features and properties available at <link will be added during post-initial-review revision>.

Methods

Genotype datasets

Release v5a of Phase 3 of The 1000 Genomes Project (accessed March 29th, 2015) provides phased genotypes at 84,801,880 genetic variants in 2,504 individuals from 26 worldwide human populations discovered using a low-coverage WGS approach [155]. During the genotype phasing, occasional positions with missing genotypes were imputed; consequently, our datasets contain no missing data. We first developed a subset of this WGS dataset in which to perform individual-level quality control prior to developing different subsets in which to evaluate the performance of the $wLOD$ method. In all subsets we applied a common set of quality-control procedures described in Pemberton *et al*. [237] to remove low-quality variants (Additional File 1: **Figure S22**).

Individual-level quality control

To independently verify the putative unrelatedness and population labeling of individuals reported by The 1000 Genomes Project Consortium, we developed a preliminary Omni dataset comprised of the 2,165,831 autosomal, 48,458 X-chromosomal, and 543 Y-chromosomal SNPs in The 1000 Genomes Project data that are present on the Illumina HumanOmni2.5-8 BeadChip (stage 1; Additional File 1: **Figure S22**). Across the 1,693 individuals for which genotypes derived using the HumanOmni2.5-8 BeadChip were also available, genotype concordance between the WGS- and BeadChip-derived genotypes lay between 0.99431 and 0.99986 (mean=0.99953, SD=0.00041). We identified intra- and inter-population pairs of individuals related closer than first cousins as well as those individuals whose reported sex or population labels were likely to be erroneous as described in Pemberton *et al*. [237]. Using these approaches, we identified six individuals whose reported sex is likely to be erroneous, 47 individuals who did not cluster genetically with other individuals sharing the same population label, and 14 intra-population and one inter-population pairs of close relatives (Additional File 2: **Table S8**).

Preparation of final datasets

Removing one individual from each intra-population relative pair, both individuals from the inter-population relative pair, and the 53 individuals whose reported sex or population labels were suspected to be erroneous (68 total individuals; Additional File 2: **Table S8**), we developed four subsets of The 1000 Genomes Project data that were restricted to the 2,436 unrelated individuals and autosomal biallelic variants (stage 2; Additional File 1: **Figure S22**).

First, we developed a WGS dataset comprised of 75,071,695 SNVs. Second, we developed a WES dataset comprised of the 1,830,512 SNVs that are present within the regions captured by the Roche Nimblegen SeqCap EZ Human Exome Library v3.0 system. Third, we developed an Omni2.5 dataset comprised of the 2,166,414 SNPs that are present on the Illumina HumanOmni2.5-8 BeadChip. Fourth, as ~96% of all markers present on the Illumina HumanOmniExpress-24 BeadChip are also present on the HumanOmni2.5-8 BeadChip, we developed an OmniExpress dataset comprised of the 676,445 SNPs in the Omni2.5 dataset that are present on the HumanOmniExpress-24 BeadChip.

Geographic distances

The geographic distance of each population from Addis Ababa, Ethiopia, was calculated as in Rosenberg *et al*. [238] with the use of waypoint routes, based on the sampling location reported by The 1000 Genomes Project [155].

Simulation of genetic datasets

For two demographic scenarios, we generated 50 independent replicates of genetic datasets using a forward-in-time process as previously described [176]. In their original approach, prior to performing the simulation steps Kardos *et al*. placed N predetermined polymorphic SNV onto the chromosome's genetic map by randomly sampling N unique genetic map positions in the range 0 to g_{max} (the user-defined genetic map length of the simulated genome), only converting genetic map positions to physical map positions based upon a fixed user-defined recombination rate to physical map distance relationship when writing the simulated datasets to file. Here, we modified their approach to instead create a non-uniform distribution of recombination rates across the simulated chromosome and allow any base pair to mutate during the simulation.

If we let g_p represent the genetic map position assigned to physical map position p, which is equal to the base pair count from the beginning of the chromosome. Based on the user-defined values for g_{max} and recombination rate θ, all values of g lie within the interval $[0, g_{max}]$ and all values of p lie within the interval $[1\,..\,(g_{max}/\theta)\times 1{,}000{,}000]$. To begin, we created a backbone of genetic and physical map positions onto which we will place all other positions, randomly drawing $(g_{max}/\theta)+1$ values of g and assigning them in increasing order to p in the range $[1\,..\,(g_{max}/\theta)]$ (i.e. every Mb). Next, we randomly chose *N* values of p to be predetermined polymorphic SNVs, and then randomly assigned each a value of g based upon the backbone interval in which it was located, again ensuring that values of g always increase as a function of p. Finally, all values of p that were not among the set of predetermined SNVs were assigned a value of g through interpolation onto the construct created by the values of p and g assigned to the predetermined SNVs. This approach created a non-uniform relationship between physical and genetic map distance along the simulated chromosome that is similar to that observed on real human chromosomes (not shown).

To extend the method of Kardos *et al*. to enable any base pair on the simulated chromosome to mutate, for each individual in each generation, the number of mutations that occur during each meiosis was drawn from a Poisson distribution with mean $\mu\times[(g_{max}/\theta)\times 1{,}000{,}000]$, where μ is mutation rate. The base pairs to be mutated were then chosen at random from all $(g_{max}/\theta)\times 1{,}000{,}000$ possible positions without replacement. Mutations were tracked and then incorporated into the genotypes of individuals in the analyzed dataset; all monomorphic positions were removed during dataset construction.

In all simulations, we set g_{max} to 325 cM, θ to 1.3 cM/Mb [239], and μ to 1.18×10^{-8} [160], and scaled θ and μ by a factor of 10 to increase genetic diversity in the final generation [240]. *N* was chosen separately for each simulated scenario such that the final number of polymorphic SNVs in the dataset (both predetermined and *de novo*) was ~750,000; *N*=725,000 for scenario 1 and *N*=650,000 for scenario 2. Because predetermined polymorphic SNVs can become fixed over the course of the simulation, their numbers in the analyzed datasets lay between 679,256-717,855 (25,788-31,503 *de novo* SNVs) for scenario 1 and between 633,582-638,675 (103,871-110,077 *de novo* SNVs) for scenario 2.

Calculation of LOD and $wLOD$ estimators

To minimize the number of variables that varied in within-dataset comparisons, we used a single set of allele frequencies when calculating $wLOD$ and LOD scores at all window sizes considered. To account for sample-size differences among populations, we used a resampling procedure to estimate the allele frequencies, sampling 100 non-missing alleles with replacement and calculating allele frequencies from these 100 alleles. As a consequence of the resampling procedure, it was possible for an individual to possess an allelic type whose frequency was estimated to be 0 in the sample of 100 alleles. SNV positions at which this scenario was encountered were treated as missing when calculating $wLOD$ and LOD scores for all windows containing the positions in individuals that had the allelic type of frequency 0.

As our datasets contained phased genotypes, in the LD correction applied in the $wLOD$ estimator (equation 3) LD was estimated with the correlation coefficient r^2 [241] using a resampling procedure to account for the possible influence of sample size on homozygosity-based LD statistics [219]. For each pair of SNPs, we randomly sampled 55 individuals—the smallest population sample size in our dataset (**Table 2**)—without replacement and the LD computation was performed using those 55 individuals. Note that we used a single set of LD estimates when calculating $wLOD$ scores at all window sizes considered.

In the recombination rate correction applied in the $wLOD$ estimator (equation 4), the genetic map position of each marker in the Omni2.5 dataset and its subsets were downloaded from the Laboratory of Computational Genetics at Rutgers University (http://compgen.rutgers.edu). The genetic map position of each marker in the WES and WGS datasets was determined by interpolation onto the Rutgers linkage-physical map [242] based on their UCSC Build hg19 physical map position.

Due to computer memory requirements for Gaussian kernel density estimation, the $wLOD$ score distributions used to determine the autozygosity score thresholds in the WGS dataset considered only twenty individuals chosen at random. Based on our investigation into the effect of sample size on score threshold (Additional File 1: **Figure S10**), we do not expect this approach to have biased our detection of ROA in the WGS dataset. All genome-wide windows were, however, considered when determining optimal window sizes in the Omni2.5, OmniExpress, and WES datasets.

Classification of ROA

We ran unsupervised Gaussian fitting of the ROA length distribution using the *mclust* package (v.5.2) [243] in *R* v.3.3.3 [244], allowing component magnitudes, means, and variances to be free parameters. BIC likelihoods with increasing number of components (*G*) were calculated using the function *mclustBIC*, while final classification under the five component model was performed using the function *Mclust*.

Genomic distribution and geographic patterns of ROA

The frequency at which each SNV was present in ROA in each population was calculated as described in Pemberton *et al*. [18]. To compare the genomic distribution of ROA across populations, we calculated mean ROA frequencies in non-overlapping 50 kb windows across all SNVs polymorphic in that population that were within the window, and excluding windows that lay within the centromere and telomeres. To evaluate the similarity of ROA frequency patterns among populations, we performed classical (metric) multidimensional scaling (MDS) separately for each ROA length class based on a matrix of ROA frequency dissimilarities between all pairs of populations, calculated as one minus the Pearson correlation coefficient (*r*) of their mean ROA frequencies across windows. We then applied MDS to this matrix using *cmdscale* in *R*.

We compared population patterns in the MDS based on ROA frequencies to an MDS based on a matrix of pairwise F_{ST} among populations calculated with our WGS dataset and the method of Hudson *et al*. [245] according to the recommendations of Bhatia *et al*. [246]. The similarity of patterns in our MDS of ROA dissimilarities and those in the MDS of F_{ST} was evaluated with the Procrustes method [247].

Relationship between ROA and genomic variables

For each ROA length class, we investigated recombination rate and haplotype-based nS_L selection scores [206] for correlations with ROA frequency across the autosomes. Population-based recombination-rate estimates were obtained from Phase 3 of The 1000 Genomes Project [155] (downloaded July14th, 2014), and nS_L values for each of the 26 populations were calculated in the WGS dataset considering only SNVs with MAF > 0.05 and normalization of unstandardized scores in 100 genome-wide frequency bins with *selscan* [248]. Comparisons between ROA frequency and recombination rate and nS_L were performed as described in Pemberton *et al*. [18] considering the mean value of each variable in non-overlapping 50 kb windows, excluding windows within the centromere and telomeres, calculated across all SNV within the window for which the variable was available. Admixed Afro-European (ASW and ACB) and Mestizo (CLM, MXL, PEL, and PUR) populations and the geographically imprecise CEU (Utah residents of Northwestern European ancestry) group were omitted from geographic analyses but were included in the scatterplots.

List of Abbreviations Used

ASW, African American; ACB, Afro-Caribbean; BEB, Bengali; bp, base-pair; CEU, European American; CDX, Dai; CHB, Northern Han; CHS, Southern Han; CLM, Colombian; cM, centimorgan; ESN, Esan; FIN, Finnish; GIH, Gujarati; GWD, Gambian; GBR, British; HGDP, Human Genome Diversity Panel; HMM, hidden Markov model; IBD, identical by descent; IBS, Iberian; Indel, insertion/deletion variant; ITU, Telugu; JPT, Japanese; kb, kilobase; KHV, Kinh; LD, linkage disequilibrium; LOD, logarithm of the odds; $wLOD$, weighted logarithm of the odds; LWK, Luhya; MAF, minor allele frequency; Mb, megabase; MSL, Mende; MXL, Mexican American; NGS, next-generation sequencing; OMIM, Online Mendelian Inheritance of Man; PEL, Peruvian; PJL, Punjabi; PUR, Puerto Rican; RAM, random access memory; ROA, regions of autozygosity; ROH, runs of homozygosity; SNP, single-nucleotide polymorphism; SNV, single-nucleotide variant; STU, Sri Lankan Tamil; TSI, Toscani; WES, whole-exome sequencing; WGS, whole-genome sequencing; YRI, Yoruban.

Declarations

Ethics approval and consent to participate

This study was approved by the University of Manitoba Health Research Ethics Board (protocol ID H2013:141).

Consent for publication

All authors read and approved the final manuscript.

Availability of data and material

The raw genetic data analyzed during the current study are available in The International Genome Sample Resource repository, http://www.internationalgenome.org/data. The software *GARLIC* implementing the new method reported by this study can be obtained at https://github.com/szpiech/garlic (v1.1.0 and later).

Competing interests

The authors declare that they have no competing financial interests.

Funding

This research was supported by an institutional start-up fund from the University of Manitoba (T.J.P.), the University of Manitoba Graduate Enhancement of Tri-Council Stipends (GETS) program (A.B.), and a Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship (A.B.). Partial support to Z.A.S was provided by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG007644 (awarded to Ryan D. Hernandez, University of California - San Francisco).

Authors' contributions

T.J.P. conceived the study. T.J.P and Z.A.S. developed the weighted likelihood estimator. T.J.P prepared the datasets and performed the genetic simulations with the assistance of Z.A.S. A.B. and M.K. performed the comparative analyses. A.B. performed the genomic analyses with the assistance of T.J.P. Z.A.S. implemented the *wLOD* method in the software *GARLIC*. A.B. and T.J.P. wrote the paper with the assistance of Z.A.S. and M.K.

Acknowledgements

The authors would like to thank Noah Rosenberg for useful comments and discussions.

Description of Additional Data Files

Additional File 1: Figures S1-22
Additional File 2: Tables S1-8

References

1. Broman KW, Weber JL: **Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain**. *Am J Hum Genet* 1999, **65**(6):1493-1500.
2. Gibson J, Morton NE, Collins A: **Extended tracts of homozygosity in outbred human populations**. *Hum Mol Genet* 2006, **15**(5):789-795.
3. Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MT, Tsai FJ, Chang CF, Wu JY, Chen YT: **Long contiguous stretches of homozygosity in the human genome**. *Hum Mutat* 2006, **27**(11):1115-1121.
4. Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK: **Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia**. *Proc Natl Acad Sci U S A* 2007, **104**(50):19942-19947.
5. The International HapMap Consortium: **A second generation human haplotype map of over 3.1**

million SNPs. *Nature* 2007, **449**(7164):851-861.

6. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A: **Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals**. *Hum Mol Genet* 2007, **16**(1):1-14.
7. Curtis D, Vine AE, Knight J: **Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations**. *Ann Hum Genet* 2008, **72**(Pt 2):261-278.
8. Jakkula E, Rehnstrom K, Varilo T, Pietilainen OP, Paunio T, Pedersen NL, Defaire U, Jarvelin MR, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L: **The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population**. *Am J Hum Genet* 2008, **83**(6):787-794.
9. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF: **Runs of homozygosity in European populations**. *Am J Hum Genet* 2008, **83**(3):359-372.
10. Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN, King KS, Nelson MR, Bustamante CD: **Global distribution of genomic diversity underscores rich complex history of continental human populations**. *Genome Res* 2009, **19**(5):795-803.
11. Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB: **Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics**. *PLoS Genet* 2009, **5**(3):e1000415.
12. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF: **Genomic runs of homozygosity record population history and consanguinity**. *PLoS One* 2010, **5**(11):e13996.
13. Nothnagel M, Lu TT, Kayser M, Krawczak M: **Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans**. *Hum Mol Genet* 2010, **19**(15):2927-2935.
14. O'Dushlaine CT, Morris D, Moskvina V, Kirov G, Consortium IS, Gill M, Corvin A, Wilson JF, Cavalleri GL: **Population structure and genome-wide patterns of variation in Ireland and Britain**. *Eur J Hum Genet* 2010, **18**(11):1248-1254.
15. Gross A, Tonjes A, Kovacs P, Veeramah KR, Ahnert P, Roshyara NR, Gieger C, Rueckert IM, Loeffler M, Stoneking M, Wichmann HE, Novembre J, Stumvoll M, Scholz M: **Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays**. *BMC Genet* 2011, **12**:67.
16. Roy-Gagnon MH, Moreau C, Bherer C, St-Onge P, Sinnett D, Laprise C, Vezina H, Labuda D: **Genomic and genealogical investigation of the French Canadian founder population structure**. *Hum Genet* 2011, **129**(5):521-531.
17. Teo SM, Ku CS, Naidoo N, Hall P, Chia KS, Salim A, Pawitan Y: **A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals**. *J Hum Genet* 2011, **56**(7):524-533.
18. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ: **Genomic patterns of homozygosity in worldwide human populations**. *Am J Hum Genet* 2012, **91**(2):275-292.
19. Teo SM, Ku CS, Salim A, Naidoo N, Chia KS, Pawitan Y: **Regions of homozygosity in three Southeast Asian populations**. *J Hum Genet* 2012, **57**(2):101-108.
20. Di Gaetano C, Fiorito G, Ortu MF, Rosa F, Guarrera S, Pardini B, Cusi D, Frau F, Barlassina C, Troffa C, Argiolas G, Zaninello R, Fresu G, Glorioso N, Piazza A, Matullo G: **Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection**. *PLoS One* 2014, **9**(3):e91237.
21. Aghakhanian F, Yunus Y, Naidu R, Jinam T, Manica A, Hoh BP, Phipps ME: **Unravelling the genetic history of Negritos and indigenous populations of Southeast Asia**. *Genome Biol Evol*

2015, **7**(5):1206-1215.

22. Karafet TM, Bulayeva KB, Bulayev OA, Gurgenova F, Omarova J, Yepiskoposyan L, Savina OV, Veeramah KR, Hammer MF: **Extensive genome-wide autozygosity in the population isolates of Daghestan**. *Eur J Hum Genet* 2015, **23**(10):1405-1412.
23. Zhai G, Zhou J, Woods MO, Green JS, Parfrey P, Rahman P, Green RC: **Genetic structure of the Newfoundland and Labrador population: founder effects modulate variability**. *Eur J Hum Genet* 2015, **24**(7):1063-1070.
24. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, Gabriel SB, Belkadi A, Boisson B, Abel L, Clark AG, Greater Middle East Variome C, Alkuraya FS, Casanova JL, Gleeson JG: **Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery**. *Nat Genet* 2016, **48**(9):1071-1076.
25. Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, Fadhlaoui-Zid K, Zalloua P, Hellenthal G, Comas D: **Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa**. *Mol Biol Evol* 2017, **34**(2):318–329.
26. Cole AM, Cox S, Jeong C, Petousi N, Aryal DR, Droma Y, Hanaoka M, Ota M, Kobayashi N, Gasparini P, Montgomery H, Robbins P, Di Rienzo A, Cavalleri GL: **Genetic structure in the Sherpa and neighboring Nepalese populations**. *BMC Genomics* 2017, **18**(1):102.
27. Gilbert E, Carmi S, Ennis S, Wilson JF, Cavalleri GL: **Genomic insights into the population structure and history of the Irish Travellers**. *Sci Rep* 2017, **7**:42187.
28. Somers M, Olde Loohuis LM, Aukes MF, Pasaniuc B, de Visser KCL, Kahn RS, Sommer IE, Ophoff RA: **A genetic population isolate in the Netherlands showing extensive haplotype sharing and long regions of homozygosity**. *Genes (Basel)* 2017, **8**(5).
29. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE *et al*: **A high-coverage genome sequence from an archaic Denisovan individual**. *Science* 2012, **338**(6104):222-226.
30. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, Sawyer S, Fu Q, Heinze A, Nickel B, Dabney J, Siebauer M, White L, Burbano HA, Renaud G, Stenzel U, Lalueza-Fox C, de la Rasilla M, Rosas A, Rudan P, Brajkovic D, Kucan Z, Gusic I, Shunkov MV, Derevianko AP, Viola B, Meyer M, Kelso J, Andres AM, Paabo S: **Patterns of coding variation in the complete exomes of three Neandertals**. *Proc Natl Acad Sci U S A* 2014, **111**(18):6666-6671.
31. Prufer 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 PL, Blanche H *et al*: **The complete genome sequence of a Neanderthal from the Altai Mountains**. *Nature* 2014, **505**(7481):43-49.
32. Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, Domboroczki L, Kovari I, Pap I, Anders A, Whittle A, Dani J, Raczky P, Higham TF, Hofreiter M, Bradley DG, Pinhasi R: **Genome flux and stasis in a five millennium transect of European prehistory**. *Nat Commun* 2014, **5**:5257.
33. Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Gotherstrom A, Reich D, Dalen L: **Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth**. *Curr Biol* 2015, **25**(10):1395-1400.
34. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prufer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M *et al*: **Great ape genetic diversity and population history**. *Nature* 2013, **499**(7459):471-475.
35. Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y,

Yngvadottir B, Cooper DN, de Manuel M, Hernandez-Rodriguez J, Lobon I, Siegismund HR, Pagani L, Quail MA, Hvilsom C, Mudakikwa A, Eichler EE, Cranfield MR, Marques-Bonet T, Tyler-Smith C, Scally A: **Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding**. *Science* 2015, **348**(6231):242-245.

36. Bertolini F, Gandolfi B, Kim ES, Haase B, Lyons LA, Rothschild MF: **Evidence of selection signatures that shape the Persian cat breed**. *Mamm Genome* 2016, **27**(3-4):144-155.
37. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, Cargill M, Auton A, Reynolds A, Elkahloun AG, Castelhano M, Mosher DS, Sutter NB, Johnson GS, Novembre J, Hubisz MJ, Siepel A, Wayne RK, Bustamante CD, Ostrander EA: **A simple genetic architecture underlies morphological variation in dogs**. *PLoS Biol* 2010, **8**(8):e1000451.
38. vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jedrzejewski W, Jedrzejewska B, Sidorovich V, Greco C, Randi E, Musiani M, Kays R, Bustamante CD, Ostrander EA, Novembre J, Wayne RK: **A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids**. *Genome Res* 2011, **21**(8):1294-1305.
39. Pilot M, Dabrowski MJ, Hayrapetyan V, Yavruyan EG, Kopaliani N, Tsingarska E, Bujalska B, Kaminski S, Bogdanowicz W: **Genetic variability of the grey wolf Canis lupus in the Caucasus in comparison with Europe and the Middle East: distinct or intermediary population?** *PLoS One* 2014, **9**(4):e93828.
40. Friedenberg SG, Meurs KM, Mackay TF: **Evaluation of artificial selection in Standard Poodles using whole-genome sequencing**. *Mamm Genome* 2016.
41. Metzger J, Pfahler S, Distl O: **Variant detection and runs of homozygosity in next generation sequencing data elucidate the genetic background of Lundehund syndrome**. *BMC Genomics* 2016, **17**(1):535.
42. Pedersen NC, Pooch AS, Liu H: **A genetic assessment of the English bulldog**. *Canine Genet Epidemiol* 2016, **3**:6.
43. Kardos M, Qvarnstrom A, Ellegren H: **Inferring Individual Inbreeding and Demographic History from Segments of Identity by Descent in Ficedula Flycatcher Genome Sequences**. *Genetics* 2017, **205**(3):1319-1334.
44. Purfield DC, Berry DP, McParland S, Bradley DG: **Runs of homozygosity and population history in cattle**. *BMC Genet* 2012, **13**:70.
45. Ai H, Huang L, Ren J: **Genetic diversity, linkage disequilibrium and selection signatures in chinese and Western pigs revealed by genome-wide SNP markers**. *PLoS One* 2013, **8**(2):e56001.
46. Ferencakovic M, Hamzic E, Gredler B, Solberg TR, Klemetsdal G, Curik I, Solkner J: **Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations**. *J Anim Breed Genet* 2013, **130**(4):286-293.
47. Herrero-Medrano JM, Megens HJ, Groenen MA, Ramis G, Bosse M, Perez-Enciso M, Crooijmans RP: **Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula**. *BMC Genet* 2013, **14**:106.
48. Pertoldi C, Purfield DC, Berg P, Jensen TH, Bach OS, Vingborg R, Kristensen TN: **Genetic characterization of a herd of the endangered Danish Jutland cattle**. *J Anim Sci* 2014, **92**(6):2372-2376.
49. Gomez-Raya L, Rodriguez C, Barragan C, Silio L: **Genomic inbreeding coefficients based on the distribution of the length of runs of homozygosity in a closed line of Iberian pigs**. *Genet Sel Evol* 2015, **47**(1):81.
50. Howard JT, Maltecca C, Haile-Mariam M, Hayes BJ, Pryce JE: **Characterizing homozygosity across United States, New Zealand and Australian Jersey cow and bull populations**. *BMC Genomics* 2015, **16**:187.
51. Marras G, Gaspa G, Sorbolini S, Dimauro C, Ajmone-Marsan P, Valentini A, Williams JL, Macciotta NP: **Analysis of runs of homozygosity and their relationship with inbreeding in five**

cattle breeds farmed in Italy. *Anim Genet* 2015, **46**(2):110-121.

52. Rodriguez-Ramilo ST, Fernandez J, Toro MA, Hernandez D, Villanueva B: **Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population**. *PLoS One* 2015, **10**(4):e0124157.
53. Zavarez LB, Utsunomiya YT, Carmo AS, Neves HH, Carvalheiro R, Ferencakovic M, Perez O'Brien AM, Curik I, Cole JB, Van Tassell CP, da Silva MV, Sonstegard TS, Solkner J, Garcia JF: **Assessment of autozygosity in Nellore cows (Bos indicus) through high-density SNP genotypes**. *Front Genet* 2015, **6**:5.
54. Iacolina L, Scandura M, Goedbloed DJ, Alexandri P, Crooijmans RP, Larson G, Archibald A, Apollonio M, Schook LB, Groenen MA, Megens HJ: **Genomic diversity and differentiation of a managed island wild boar population**. *Heredity (Edinb)* 2015, **116**(1):60-67.
55. Mastrangelo S, Tolone M, Di Gerlando R, Fontanesi L, Sardina MT, Portolano B: **Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds**. *Animal* 2016, **10**(5):746-754.
56. Burren A, Neuditschko M, Signer-Hasler H, Frischknecht M, Reber I, Menzi F, Drogemuller C, Flury C: **Genetic diversity analyses reveal first insights into breed-specific selection signatures within Swiss goat breeds**. *Anim Genet* 2016, **47**(6):727-739.
57. Mastrangelo S, Portolano B, Di Gerlando R, Ciampolini R, Tolone M, Sardina MT, International Sheep Genomics C: **Genome-wide analysis in endangered populations: a case study in Barbaresca sheep**. *Animal* 2017:1-10.
58. Msalya G, Kim ES, Laisser EL, Kipanyula MJ, Karimuribo ED, Kusiluka LJ, Chenyambuga SW, Rothschild MF: **Determination of Genetic Structure and Signatures of Selection in Three Strains of Tanzania Shorthorn Zebu, Boran and Friesian Cattle by Genome-Wide SNP Analyses**. *PLoS One* 2017, **12**(1):e0171088.
59. Grossi DA, Jafarikia M, Brito LF, Buzanskas ME, Sargolzaei M, Schenkel FS: **Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs**. *BMC Genet* 2017, **18**(1):6.
60. Brito LF, Kijas JW, Ventura RV, Sargolzaei M, Porto-Neto LR, Canovas A, Feng Z, Jafarikia M, Schenkel FS: **Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers**. *BMC Genomics* 2017, **18**(1):229.
61. Purfield DC, McParland S, Wall E, Berry DP: **The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds**. *PLoS One* 2017, **12**(5):e0176780.
62. Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LA, Schook LB, Crooijmans RP, Groenen MA: **Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape**. *PLoS Genet* 2012, **8**(11):e1003100.
63. Kim ES, Cole JB, Huson H, Wiggans GR, Van Tassell CP, Crooker BA, Liu G, Da Y, Sonstegard TS: **Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle**. *PLoS One* 2013, **8**(11):e80813.
64. Silio L, Rodriguez MC, Fernandez A, Barragan C, Benitez R, Ovilo C, Fernandez AI: **Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics**. *J Anim Breed Genet* 2013, **130**(5):349-360.
65. Pryce JE, Haile-Mariam M, Goddard ME, Hayes BJ: **Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle**. *Genet Sel Evol* 2014, **46**:71.
66. Kim ES, Sonstegard TS, Rothschild MF: **Recent artificial selection in U.S. Jersey cattle impacts autozygosity levels of specific genomic regions**. *BMC Genomics* 2015, **16**:302.
67. Kim ES, Sonstegard TS, Van Tassell CP, Wiggans G, Rothschild MF: **The relationship between runs of homozygosity and inbreeding in Jersey cattle under selection**. *PLoS One* 2015, **10**(7):e0129967.
68. Metzger J, Karwath M, Tonda R, Beltran S, Agueda L, Gut M, Gut IG, Distl O: **Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses**. *BMC Genomics* 2015, **16**(1):764.

69. Saura M, Fernandez A, Varona L, Fernandez AI, de Cara MA, Barragan C, Villanueva B: **Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data**. *Genet Sel Evol* 2015, **47**:1.
70. Zhang Q, Guldbrandtsen B, Bosse M, Lund MS, Sahana G: **Runs of homozygosity and distribution of functional variants in the cattle genome**. *BMC Genomics* 2015, **16**:542.
71. Zanella R, Peixoto JO, Cardoso FF, Cardoso LL, Biegelmeyer P, Cantao ME, Otaviano A, Freitas MS, Caetano AR, Ledur MC: **Genetic diversity analysis of two commercial breeds of pigs using genomic and pedigree data**. *Genet Sel Evol* 2016, **48**(1):24.
72. Andrea T, Bertolini F, Pagnacco G, Pilla F, Ajmone-Marsan P, Rothschild MF, Crepaldi P, Italian Goat C: **The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions**. *Mamm Genome* 2017.
73. Ferencakovic M, Solkner J, Kaps M, Curik I: **Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population**. *J Dairy Sci* 2017.
74. Luan T, Yu X, Dolezal M, Bagnato A, Meuwissen TH: **Genomic prediction based on runs of homozygosity**. *Genet Sel Evol* 2014, **46**:64.
75. Gurgul A, Szmatola T, Topolski P, Jasielczuk I, Zukowski K, Bugno-Poniewierska M: **The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle**. *J Appl Genet* 2016.
76. Meszaros G, Boison SA, Perez O'Brien AM, Ferencakovic M, Curik I, Da Silva MV, Utsunomiya YT, Garcia JF, Solkner J: **Genomic analysis for managing small and endangered populations: a case study in Tyrol Grey cattle**. *Front Genet* 2015, **6**:173.
77. Howard JT, Tiezzi F, Huang Y, Gray KA, Maltecca C: **Characterization and management of long runs of homozygosity in parental nucleus lines and their associated crossbred progeny**. *Genet Sel Evol* 2016, **48**(1):91.
78. Maatouk F, Laamiri D, Argoubi K, Ghedira H: **Dental manifestations of inbreeding**. *J Clin Pediatr Dent* 1995, **19**(4):305-306.
79. Zlotogora J: **Genetic disorders among Palestinian Arabs: 1. Effects of consanguinity**. *Am J Med Genet* 1997, **68**(4):472-475.
80. Stoll C, Alembik Y, Roth MP, Dott B: **Parental consanguinity as a cause for increased incidence of births defects in a study of 238,942 consecutive births**. *Ann Genet* 1999, **42**(3):133-139.
81. Modell B, Darr A: **Science and society: genetic counselling and customary consanguineous marriage**. *Nat Rev Genet* 2002, **3**(3):225-229.
82. Habeb AM, Flanagan SE, Deeb A, Al-Alwan I, Alawneh H, Balafrej AAL, Mutair A, Hattersley AT, Hussain K, Ellard S: **Permanent neonatal diabetes: different aetiology in Arabs compared to Europeans**. *Arch Dis Child* 2012, **97**(8):721-723.
83. Morton NE: **Effect of inbreeding on IQ and mental retardation**. *Proc Natl Acad Sci U S A* 1978, **75**(8):3906-3908.
84. Shami SA, Qaisar R, Bittles AH: **Consanguinity and adult morbidity in Pakistan**. *Lancet* 1991, **338**(8772):954.
85. Rudan I: **Inbreeding and cancer incidence in human isolates**. *Hum Biol* 1999, **71**(2):173-187.
86. Mani A, Meraji SM, Houshyar R, Radhakrishnan J, Ahangar M, Rezaie TM, Taghavinejad MA, Broumand B, Zhao H, Nelson-Williams C, Lifton RP: **Finding genetic contributions to sporadic disease: a recessive locus at 12q24 commonly contributes to patent ductus arteriosus**. *Proc Natl Acad Sci U S A* 2002, **99**(23):15054-15059.
87. Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, Rudan P: **Inbreeding and risk of late onset complex disease**. *J Med Genet* 2003, **40**(12):925-932.
88. Kanaan ZM, Mahfouz R, Tamim H: **The prevalence of consanguineous marriages in an underserved area in Lebanon and its association with congenital anomalies**. *Genet Test* 2008, **12**(3):367-372.
89. Shieh JTC, Bittles AH, Hudgins L: **Consanguinity and the risk of congenital heart disease**. *Am J Med Genet A* 2012, **158A**(5):1236-1241.

90. Fareed M, Afzal M: **Increased cardiovascular risks associated with familial inbreeding: a population-based study of adolescent cohort**. *Ann Epidemiol* 2016, **26**(4):283-292.
91. el-Orfi AHA, Singh M, Giasuddin ASM: **Conjugal leprosy among Libyan patients**. *Dermatology* 1998, **196**(2):271-272.
92. Lyons EJ, Frodsham AJ, Zhang L, Hill AV, Amos W: **Consanguinity and susceptibility to infectious diseases in humans**. *Biol Lett* 2009, **5**(4):574-576.
93. Lyons EJ, Amos W, Berkley JA, Mwangi I, Shafi M, Williams TN, Newton CR, Peshu N, Marsh K, Scott JAG, Hill AVS: **Homozygosity and risk of childhood death due to invasive bacterial disease**. *BMC Med Genet* 2009, **10**:55.
94. Garrod AE: **The incidence of alkaptonuria: a study in chemical individuality**. *Lancet* 1902, **ii**:1616-1620.
95. Pemberton TJ, Rosenberg NA: **Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective**. *Hum Hered* 2014, **77**(1-4):37-48.
96. Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zollner S, Rosenberg NA, Li JZ: **Long runs of homozygosity are enriched for deleterious variation**. *Am J Hum Genet* 2013, **93**(1):90-102.
97. Mezzavilla M, Vozzi D, Badii R, Alkowari MK, Abdulhadi K, Girotto G, Gasparini P: **Increased rate of deleterious variants in long runs of homozygosity of an inbred population from Qatar**. *Hum Hered* 2015, **79**(1):14-19.
98. McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko T, Corre T, Davies G, Kaakinen M, Lyytikainen LP, Kristiansson K, Havulinna AS, Gogele M, Vitart V, Tenesa A, Aulchenko Y, Hayward C, Johansson A, Boban M, Ulivi S, Robino A, Boraska V, Igl W, Wild SH, Zgaga L, Amin N, Theodoratou E, Polasek O, Girotto G, Lopez LM *et al*: **Evidence of inbreeding depression on human height**. *PLoS Genet* 2012, **8**(7):e1002655.
99. Fareed M, Afzal M: **Evidence of inbreeding depression on height, weight, and body mass index: A population-based child cohort study**. *Am J Hum Biol* 2014, **26**(6):784-795.
100. Verweij KJ, Abdellaoui A, Veijola J, Sebert S, Koiranen M, Keller MC, Jarvelin MR, Zietsch BP: **The association of genotype-based inbreeding coefficient with a range of physical and psychological human traits**. *PLoS One* 2014, **9**(7):e103102.
101. Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, Nutile T, Jackson AU, Schurmann C, Smith AV, Zhang W, Okada Y, Stancakova A, Faul JD, Zhao W, Bartz TM, Concas MP, Franceschini N, Enroth S, Vitart V, Trompet S, Guo X, Chasman DI, O'Connel JR, Corre T, Nongmaithem SS, Chen Y, Mangino M, Ruggiero D, Traglia M, Farmaki AE *et al*: **Directional dominance on stature and cognition in diverse human populations**. *Nature* 2015, **523**(7561):459-462.
102. Power RA, Nagoshi C, DeFries JC, Plomin R, Wellcome Trust Case Control C: **Genome-wide estimates of inbreeding in unrelated individuals and their association with cognitive ability**. *Eur J Hum Genet* 2014, **22**(3):386-390.
103. Howrigan DP, Simonson MA, Davies G, Harris SE, Tenesa A, Starr JM, Liewald DC, Deary IJ, McRae A, Wright MJ, Montgomery GW, Hansell N, Martin NG, Payton A, Horan M, Ollier WE, Abdellaoui A, Boomsma DI, DeRosse P, Knowles EE, Glahn DC, Djurovic S, Melle I, Andreassen OA, Christoforou A, Steen VM, Hellard SL, Sundet K, Reinvang I, Espeseth T *et al*: **Genome-wide autozygosity is associated with lower general cognitive ability**. *Mol Psychiatry* 2015, **21**(6):837-843.
104. Krieger H: **Inbreeding effects on metrical traits in Northeastern Brazil**. *Am J Hum Genet* 1969, **21**(6):537-546.
105. Martin AO, Kurczynski TW, Steinberg AG: **Familial studies of medical and anthropometric variables in a human isolate**. *Am J Hum Genet* 1973, **25**(6):581-593.
106. Hurwich BJ, Nubani N: **Blood pressures in a highly inbred community--Abu Ghosh, Israel. 1. Original survey**. *Isr J Med Sci* 1978, **14**(9):962-969.
107. Halberstein RA: **Blood pressure in the Caribbean**. *Hum Biol* 1999, **71**(4):659-684.
108. Saleh EA, Mahfouz AA, Tayel KY, Naguib MK, Bin-al-Shaikh NM: **Hypertension and its determinants among primary-school children in Kuwait: an epidemiological study**. *East*

Mediterr Health J 2000, **6**(2-3):333-337.

109. Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, Rudan P: **Inbreeding and the genetic complexity of human hypertension**. *Genetics* 2003, **163**(3):1011-1021.
110. Badaruddoza: **Inbreeding effects on metrical phenotypes among North Indian Children**. *Coll Antropol* 2004, **28 Suppl 2**:311-319.
111. Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, Kolcic I, Weber JL, Hastie ND, Rudan P, Wright AF: **Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits**. *Hum Mol Genet* 2007, **16**(2):233-241.
112. Simpson JL, Martin AO, Elias S, Sarto GE, Dunn JK: **Cancers of the breast and female genital system: search for recessive genetic factors through analysis of human isolate**. *Am J Obstet Gynecol* 1981, **141**(6):629-636.
113. Lebel RR, Gallagher WB: **Wisconsin consanguinity studies. II: Familial adenocarcinomatosis**. *Am J Med Genet* 1989, **33**(1):1-6.
114. Bacolod MD, Schemmann GS, Wang S, Shattock R, Giardina SF, Zeng Z, Shia J, Stengel RF, Gerry N, Hoh J, Kirchhoff T, Gold B, Christman MF, Offit K, Gerald WL, Notterman DA, Ott J, Paty PB, Barany F: **The signatures of autozygosity among patients with colorectal cancer**. *Cancer Res* 2008, **68**(8):2610-2621.
115. Thomsen H, Inacio da Silva Filho M, Fuchs M, Ponader S, Pogge von Strandmann E, Eisele L, Herms S, Hoffmann P, Engert A, Hemminki K, Forsti A: **Evidence of Inbreeding in Hodgkin Lymphoma**. *PLoS One* 2016, **11**(4):e0154259.
116. Thomsen H, Chen B, Figlioli G, Elisei R, Romei C, Cipollini M, Cristaudo A, Bambi F, Hoffmann P, Herms S, Landi S, Hemminki K, Gemignani F, Forsti A: **Runs of homozygosity and inbreeding in thyroid cancer**. *BMC Cancer* 2016, **16**(1):227.
117. Puzyrev VP, Lemza SV, Nazarenko LP, Panphilov VI: **Influence of genetic and demographic factors on etiology and pathogenesis of chronic disease in north Siberian aborigines**. *Arctic Med Res* 1992, **51**(3):136-142.
118. Ismail J, Jafar TH, Jafary FH, White F, Faruqui AM, Chaturvedi N: **Risk factors for non-fatal myocardial infarction in young South Asian adults**. *Heart* 2004, **90**(3):259-263.
119. Christofidou P, Nelson CP, Nikpay M, Qu L, Li M, Loley C, Debiec R, Braund PS, Denniff M, Charchar FJ, Arjo AR, Tregouet DA, Goodall AH, Cambien F, Ouwehand WH, Roberts R, Schunkert H, Hengstenberg C, Reilly MP, Erdmann J, McPherson R, Konig IR, Thompson JR, Samani NJ, Tomaszewski M: **Runs of Homozygosity: Association with Coronary Artery Disease and Gene Expression in Monocytes and Macrophages**. *Am J Hum Genet* 2015, **97**(2):228-237.
120. McLaughlin RL, Kenna KP, Vajda A, Heverin M, Byrne S, Donaghy CG, Cronin S, Bradley DG, Hardiman O: **Homozygosity mapping in an Irish ALS case-control cohort describes local demographic phenomena and points towards potential recessive risk loci**. *Genomics* 2015, **105**(4):237-241.
121. Keller MC, Miller G: **Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best?** *Behav Brain Sci* 2006, **29**(4):385-404; discussion 405-352.
122. Keller MC, Simonson MA, Ripke S, Neale BM, Gejman PV, Howrigan DP, Lee SH, Lencz T, Levinson DF, Sullivan PF, Schizophrenia Psychiatric Genome-Wide Association Study C: **Runs of homozygosity implicate autozygosity as a schizophrenia risk factor**. *PLoS Genet* 2012, **8**(4):e1002656.
123. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA: **Mapping complex disease loci in whole-genome association studies**. *Nature* 2004, **429**(6990):446-452.
124. Freimer N, Sabatti C: **The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology**. *Nat Genet* 2004, **36**(10):1045-1051.

125. Pritchard JK: **Are rare variants responsible for susceptibility to complex diseases?** *Am J Hum Genet* 2001, **69**(1):124-137.
126. Pritchard JK, Cox NJ: **The allelic architecture of human disease genes: common disease-common variant...or not?** *Hum Mol Genet* 2002, **11**(20):2417-2423.
127. Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H: **A polygenic basis for late-onset disease**. *Trends Genet* 2003, **19**(2):97-106.
128. Reich DE, Lander ES: **On the allelic spectrum of human disease**. *Trends Genet* 2001, **17**(9):502-510.
129. Yang TL, Guo Y, Zhang LS, Tian Q, Yan H, Papasian CJ, Recker RR, Deng HW: **Runs of homozygosity identify a recessive locus 12q21.31 for human adult height**. *J Clin Endocrinol Metab* 2010, **95**(8):3777-3782.
130. Yang HC, Chang LC, Liang YJ, Lin CH, Wang PL: **A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human major histocompatibility complex**. *PLoS One* 2012, **7**(4):e34840.
131. Simon-Sanchez J, Kilarski LL, Nalls MA, Martinez M, Schulte C, Holmans P, International Parkinson's Disease Genomics C, Wellcome Trust Case Control C, Gasser T, Hardy J, Singleton AB, Wood NW, Brice A, Heutink P, Williams N, Morris HR: **Cooperative genome-wide analysis shows increased homozygosity in early onset Parkinson's disease**. *PLoS One* 2012, **7**(3):e28787.
132. Ghani M, Sato C, Lee JH, Reitz C, Moreno D, Mayeux R, St George-Hyslop P, Rogaeva E: **Evidence of recessive Alzheimer disease loci in a Caribbean Hispanic data set: genome-wide survey of runs of homozygosity**. *JAMA Neurol* 2013, **70**(10):1261-1267.
133. Ghani M, Reitz C, Cheng R, Vardarajan BN, Jun G, Sato C, Naj A, Rajbhandary R, Wang LS, Valladares O, Lin CF, Larson EB, Graff-Radford NR, Evans D, De Jager PL, Crane PK, Buxbaum JD, Murrell JR, Raj T, Ertekin-Taner N, Logue M, Baldwin CT, Green RC, Barnes LL, Cantwell LB, Fallin MD, Go RC, Griffith PA, Obisesan TO, Manly JJ *et al*: **Association of Long Runs of Homozygosity With Alzheimer Disease Among African American Individuals**. *JAMA Neurol* 2015, **72**(11):1313-1323.
134. Mukherjee S, Guha S, Ikeda M, Iwata N, Malhotra AK, Pe'er I, Darvasi A, Lencz T: **Excess of homozygosity in the major histocompatibility complex in schizophrenia**. *Hum Mol Genet* 2014, **23**(22):6088-6095.
135. Sud A, Cooke R, Swerdlow AJ, Houlston RS: **Genome-wide homozygosity signature and risk of Hodgkin lymphoma**. *Sci Rep* 2015, **5**:14315.
136. Lander ES, Botstein D: **Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children**. *Science* 1987, **236**(4808):1567-1570.
137. Ben Hamida C, Doerflinger N, Belal S, Linder C, Reutenauer L, Dib C, Gyapay G, Vignal A, Le Paslier D, Cohen D, et al.: **Localization of Friedreich ataxia phenotype with selective vitamin E deficiency to chromosome 8q by homozygosity mapping**. *Nat Genet* 1993, **5**(2):195-200.
138. Kwitek-Black AE, Carmi R, Duyk GM, Buetow KH, Elbedour K, Parvari R, Yandava CN, Stone EM, Sheffield VC: **Linkage of Bardet-Biedl syndrome to chromosome 16q and evidence for non-allelic genetic heterogeneity**. *Nat Genet* 1993, **5**(4):392-396.
139. Pollak MR, Chou YH, Cerda JJ, Steinmann B, La Du BN, Seidman JG, Seidman CE: **Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2**. *Nat Genet* 1993, **5**(2):201-204.
140. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease**. *Nat Genet* 2003, **33 Suppl**:228-237.
141. Hildebrandt F, Heeringa SF, Ruschendorf F, Attanasio M, Nurnberg G, Becker C, Seelow D, Huebner N, Chernin G, Vlangos CN, Zhou W, O'Toole JF, Hoskins BE, Wolf MT, Hinkes BG, Chaib H, Ashraf S, Schoeb DS, Ovunc B, Allen SJ, Vega-Warner V, Wise E, Harville HM, Lyons RH, Washburn J, Macdonald J, Nurnberg P, Otto EA: **A systematic approach to mapping recessive disease genes in individuals from outbred populations**. *PLoS Genet* 2009, **5**(1):e1000353.

142. Collin RW, van den Born LI, Klevering BJ, de Castro-Miro M, Littink KW, Arimadyo K, Azam M, Yazar V, Zonneveld MN, Paun CC, Siemiatkowska AM, Strom TM, Hehir-Kwa JY, Kroes HY, de Faber JT, van Schooneveld MJ, Heckenlively JR, Hoyng CB, den Hollander AI, Cremers FP: **High-resolution homozygosity mapping is a powerful tool to detect novel mutations causative of autosomal recessive RP in the Dutch population**. *Invest Ophthalmol Vis Sci* 2011, **52**(5):2227-2239.
143. Hagiwara K, Morino H, Shiihara J, Tanaka T, Miyazawa H, Suzuki T, Kohda M, Okazaki Y, Seyama K, Kawakami H: **Homozygosity mapping on homozygosity haplotype analysis to detect recessive disease-causing genes from a small number of unrelated, outbred patients**. *PLoS One* 2011, **6**(9):e25059.
144. Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, LifeLines Cohort S, Esko T, Milani L, Magi R, Metspalu A, Hill WG, Weir BS, Goddard ME, Visscher PM, Yang J: **Dominance genetic variation contributes little to the missing heritability for human complex traits**. *Am J Hum Genet* 2015, **96**(3):377-385.
145. Gibbs JR, Singleton A: **Application of genome-wide single nucleotide polymorphism typing: simple association and beyond**. *PLoS Genet* 2006, **2**(10):e150.
146. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.
147. Browning SR, Browning BL: **High-resolution detection of identity by descent in unrelated individuals**. *Am J Hum Genet* 2010, **86**(4):526-539.
148. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I: **Whole population, genome-wide mapping of hidden relatedness**. *Genome Res* 2009, **19**(2):318-326.
149. Howrigan DP, Simonson MA, Keller MC: **Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms**. *BMC Genomics* 2011, **12**:460.
150. Wang S, Haynes C, Barany F, Ott J: **Genome-wide autozygosity mapping in human populations**. *Genet Epidemiol* 2009, **33**(2):172-180.
151. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
152. Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, Risch N: **Estimating genotype error rates from high-coverage next-generation sequence data**. *Genome Res* 2014, **24**(11):1734-1739.
153. Clark AG: **The size distribution of homozygous segments in the human genome**. *Am J Hum Genet* 1999, **65**(6):1489-1492.
154. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R: **BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data**. *Bioinformatics* 2016, **32**(11):1749-1751.
155. The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
156. Chen H, Patterson N, Reich D: **Population differentiation as a test for selective sweeps**. *Genome Res* 2010, **20**(3):393-402.
157. Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R: **Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium**. *Genet Epidemiol* 2009, **33**(3):266-274.
158. Keller MC, Visscher PM, Goddard ME: **Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data**. *Genetics* 2011, **189**(1):237-249.
159. MacLeod IM, Larkin DM, Lewin HA, Hayes BJ, Goddard ME: **Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors**. *Mol Biol*

Evol 2013, **30**(9):2209-2223.
160. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P: **Variation in genome-wide mutation rates within and between human families**. *Nat Genet* 2011, **43**(7):712-714.
161. Schiffels S, Durbin R: **Inferring human population size and separation history from multiple genome sequences**. *Nat Genet* 2014, **46**(8):919-925.
162. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB: **Genotype, haplotype and copy-number variation in worldwide human populations**. *Nature* 2008, **451**(7181):998-1003.
163. Pemberton TJ, Jakobsson M, Conrad DF, Coop G, Wall JD, Pritchard JK, Patel PI, Rosenberg NA: **Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India**. *Ann Hum Genet* 2008, **72**(Pt 4):535-546.
164. Prugnolle F, Manica A, Balloux F: **Geography predicts neutral genetic diversity of human populations**. *Curr Biol* 2005, **15**(5):R159-160.
165. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: **Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa**. *Proc Natl Acad Sci U S A* 2005, **102**(44):15942-15947.
166. DeGiorgio M, Jakobsson M, Rosenberg NA: **Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa**. *Proc Natl Acad Sci U S A* 2009, **106**(38):16057-16062.
167. Pemberton TJ, DeGiorgio M, Rosenberg NA: **Population structure in a comprehensive genomic data set on human microsatellite variation**. *G3 (Bethesda)* 2013, **3**(5):891-907.
168. Xu S, Huang W, Wang H, He Y, Wang Y, Wang Y, Qian J, Xiong M, Jin L: **Dissecting linkage disequilibrium in African-American genomes: roles of markers and individuals**. *Mol Biol Evol* 2007, **24**(9):2049-2058.
169. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, Contreras A, Balam-Ortiz E, del Bosque-Plata L, Velazquez-Fernandez D, Lara C, Goya R, Hernandez-Lemus E, Davila C, Barrientos E, March S, Jimenez-Sanchez G: **Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico**. *Proc Natl Acad Sci U S A* 2009, **106**(21):8611-8616.
170. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H: **Genome-wide patterns of population structure and admixture among Hispanic/Latino populations**. *Proc Natl Acad Sci U S A* 2010, **107 Suppl 2**:8954-8961.
171. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, Camrena B, Nicolini H, Klitz W, Barrantes R, Molina JA, Freimer NB, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Dipierri JE, Alfaro EL, Bailliet G, Bianchi NO, Llop E, Rothhammer F, Excoffier L, Ruiz-Linares A: **Geographic patterns of genome admixture in Latin American Mestizos**. *PLoS Genet* 2008, **4**(3):e1000037.
172. Wu L: **Investigation of Consanguineous Marriages among 30 Chinese Ethnic Groups**. *Hered Dis* 1987, **4**:163-166.
173. Bittles AH: **Consangunity in context**. Cambridge , UK: Cambridge University Press; 2012.
174. Zhusheng W: **Dai**. In. Encyclopedia.com; 1996.
175. Gadgil M, Joshi NV, Prasad UVS, Manoharan S, Patil S: **Peopling of India**. In: *The Indian Human Heritage*. Edited by Balasubramanian D, Rao NA. Hyderabad, India: Universities Press; 1998: 100-129.
176. Kardos M, Luikart G, Allendorf FW: **Measuring individual inbreeding in the age of genomics:**

marker-based measures are better than pedigrees. *Heredity (Edinb)* 2015, **115**(1):63-72.

177. Fenner JN: **Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies**. *Am J Phys Anthropol* 2005, **128**(2):415-423.
178. Szpiech ZA, Blant A, Pemberton TJ: ***GARLIC*: Genomic Autozygosity Regions Likelihood-based Inference and Classification**. *Bioinformatics* 2017, **doi**:10.1093/bioinformatics/btx1102.
179. Browning BL, Browning SR: **Improving the accuracy and efficiency of identity-by-descent detection in population data**. *Genetics* 2013, **194**(2):459-471.
180. Thomas A, Skolnick MH, Lewis CM: **Genomic mismatch scanning in pedigrees**. *IMA J Math Appl Med Biol* 1994, **11**(1):1-16.
181. Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA: **Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays**. *Ann Hum Genet* 2008, **72**(Pt 2):279-287.
182. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.
183. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: **A worldwide survey of haplotype variation and linkage disequilibrium in the human genome**. *Nat Genet* 2006, **38**(11):1251-1260.
184. Huang L, Jakobsson M, Pemberton TJ, Ibrahim M, Nyambo T, Omar S, Pritchard JK, Tishkoff SA, Rosenberg NA: **Haplotype variation and genotype imputation in African populations**. *Genet Epidemiol* 2011, **35**(8):766-780.
185. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation**. *Science* 2008, **319**(5866):1100-1104.
186. Gazal S, Sahbatou M, Babron MC, Genin E, Leutenegger AL: **High level of inbreeding in final phase of 1000 Genomes Project**. *Sci Rep* 2015, **5**:17453.
187. Shen Z, Duan L, Yang H, Yuan L, Huang Y, Li L, Xu B: **Genetic variation of 17 STR loci in Dai population in mainland China**. *Forensic Sci Int Genet* 2015, **19**:37-38.
188. Sun H, Zhou C, Huang X, Lin K, Shi L, Yu L, Liu S, Chu J, Yang Z: **Autosomal STRs provide genetic evidence for the hypothesis that Tai people originate from southern China**. *PLoS One* 2013, **8**(4):e60822.
189. Li YC, Huang W, Tian JY, Chen XQ, Kong QP: **Exploring the maternal history of the Tai people**. *J Hum Genet* 2016, **61**(8):721-729.
190. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M: **Natural selection on genes that underlie human disease susceptibility**. *Curr Biol* 2008, **18**(12):883-889.
191. Berg JS, Adams M, Nassar N, Bizon C, Lee K, Schmitt CP, Wilhelmsen KC, Evans JP: **An informatics approach to analyzing the incidentalome**. *Genet Med* 2013, **15**(1):36-44.
192. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A *et al*: **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature* 2016, **536**(7616):285-291.
193. Itan Y, Jones BL, Ingram CJ, Swallow DM, Thomas MG: **A worldwide correlation of lactase persistence phenotype and genotypes**. *BMC Evol Biol* 2010, **10**:36.
194. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG: **Evolution of lactase persistence: an example of human niche construction**. *Philos Trans R Soc Lond B Biol Sci* 2011, **366**(1566):863-877.
195. Smith GD, Lawlor DA, Timpson NJ, Baban J, Kiessling M, Day IN, Ebrahim S: **Lactase persistence-related genetic variant: population substructure and health outcomes**. *Eur J Hum*

Genet 2009, **17**(3):357-367.

196. Ingram CJ, Mulcare CA, Itan Y, Thomas MG, Swallow DM: **Lactose digestion and the evolutionary genetics of lactase persistence**. *Hum Genet* 2009, **124**(6):579-591.
197. Simoons FJ: **Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis**. *Am J Dig Dis* 1970, **15**(8):695-710.
198. McCracken RD: **Lactase Deficiency: An Example of Dietary Evolution**. *Curr Anthropol* 1971, **12**(4/5):479-517.
199. Kretchmer N: **Lactose and lactase**. *Sci Am* 1972, **227**(4):71-78.
200. Wang YG, Yan YS, Xu JJ, Du RF, Flatz SD, Kuhnau W, Flatz G: **Prevalence of primary adult lactose malabsorption in three populations of northern China**. *Hum Genet* 1984, **67**(1):103-106.
201. Gallego Romero I, Basu Mallick C, Liebert A, Crivellaro F, Chaubey G, Itan Y, Metspalu M, Eaaswarkhanth M, Pitchappan R, Villems R, Reich D, Singh L, Thangaraj K, Thomas MG, Swallow DM, Mirazon Lahr M, Kivisild T: **Herders of Indian and European cattle share their predominant allele for lactase persistence**. *Mol Biol Evol* 2012, **29**(1):249-260.
202. Woteki CE, Weser E, Young EA: **Lactose malabsorption in Mexican-American children**. *Am J Clin Nutr* 1976, **29**(1):19-24.
203. Sahi T: **Genetics and epidemiology of adult-type hypolactasia**. *Scand J Gastroenterol Suppl* 1994, **202**:7-20.
204. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I: **Identification of a variant associated with adult-type hypolactasia**. *Nat Genet* 2002, **30**(2):233-237.
205. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P: **Convergent adaptation of human lactase persistence in Africa and Europe**. *Nat Genet* 2007, **39**(1):31-40.
206. Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R: **On detecting incomplete soft or hard selective sweeps using haplotype structure**. *Mol Biol Evol* 2014, **31**(5):1275-1291.
207. McVean G: **A genealogical interpretation of principal components analysis**. *PLoS Genet* 2009, **5**(10):e1000686.
208. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis**. *PLoS Genet* 2006, **2**(12):e190.
209. Pemberton TJ, Li FY, Hanson EK, Mehta NU, Choi S, Ballantyne J, Belmont JW, Rosenberg NA, Tyler-Smith C, Patel PI: **Impact of restricted marital practices on genetic variation in an endogamous Gujarati group**. *Am J Phys Anthropol* 2012, **149**(1):92-103.
210. Hodges JL, Lehmann EL: **The efficiency of some nonparametric competitors of the *t*-test**. *Ann Math Statist* 1956, **27**(2):324-335.
211. Chen G, Bentley A, Adeyemo A, Shriner D, Zhou J, Doumatey A, Huang H, Ramos E, Erdos M, Gerry N, Herbert A, Christman M, Rotimi C: **Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans**. *Hum Mol Genet* 2012, **21**(20):4530-4536.
212. Maeda N, Shimomura I, Kishida K, Nishizawa H, Matsuda M, Nagaretani H, Furuyama N, Kondo H, Takahashi M, Arita Y, Komuro R, Ouchi N, Kihara S, Tochino Y, Okutomi K, Horie M, Takeda S, Aoyama T, Funahashi T, Matsuzawa Y: **Diet-induced insulin resistance in mice lacking adiponectin/ACRP30**. *Nat Med* 2002, **8**(7):731-737.
213. Diez JJ, Iglesias P: **The role of the novel adipocyte-derived hormone adiponectin in human disease**. *Eur J Endocrinol* 2003, **148**(3):293-300.
214. Sim X, Ong RT, Suo C, Tay WT, Liu J, Ng DP, Boehnke M, Chia KS, Wong TY, Seielstad M, Teo YY, Tai ES: **Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia**. *PLoS Genet* 2011, **7**(4):e1001363.
215. Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, Ng SM, Erazo M, Ostrer H, Abraham C, Abreu MT, Atzmon G, Barzilai N, Brant SR, Bressman S, Burns ER, Chowers Y, Clark LN,

Darvasi A, Doheny D, Duerr RH, Eliakim R, Giladi N, Gregersen PK, Hakonarson H, Jones MR, Marder K, McGovern DP, Mulle J, Orr-Urtreger A, Proctor DD *et al*: **A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci**. *PLoS Genet* 2012, **8**(3):e1002559.

216. Walker SM, Downes CP, Leslie NR: **TPIP: a novel phosphoinositide 3-phosphatase**. *Biochem J* 2001, **360**(Pt 2):277-283.
217. Clifford RJ, Zhang J, Meerzaman DM, Lyu MS, Hu Y, Cultraro CM, Finney RP, Kelley JM, Efroni S, Greenblum SI, Nguyen CV, Rowe WL, Sharma S, Wu G, Yan C, Zhang H, Chung YH, Kim JA, Park NH, Song IH, Buetow KH: **Genetic variations at loci involved in the immune response are risk factors for hepatocellular carcinoma**. *Hepatology* 2010, **52**(6):2034-2043.
218. Zuo L, Wang T, Lin X, Wang J, Tan Y, Wang X, Yu X, Luo X: **Sex difference of autosomal alleles in populations of European and African descent**. *Genes Genomics* 2015, **37**(12):1007-1016.
219. Rosenberg NA, Blum MG: **Sampling properties of homozygosity-based statistics for linkage disequilibrium**. *Math Biosci* 2007, **208**(1):33-47.
220. Fung T, Keenan K: **Confidence intervals for population allele frequencies: the general case of sampling from a finite diploid population of any size**. *PLoS One* 2014, **9**(1):e85925.
221. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A *et al*: **A human genome diversity cell line panel**. *Science* 2002, **296**(5566):261-262.
222. Cavalli-Sforza LL: **The Human Genome Diversity Project: past, present and future**. *Nat Rev Genet* 2005, **6**(4):333-340.
223. The International HapMap 3 Consortium: **Integrating common and rare genetic variation in diverse human populations**. *Nature* 2010, **467**(7311):52-58.
224. Baerenwald DA, Bonnefond A, Bouatia-Naji N, Flemming BP, Umunakwe OC, Oeser JK, Pound LD, Conley NL, Cauchi S, Lobbens S, Eury E, Balkau B, Lantieri O, Investigators M, Dadi PK, Jacobson DA, Froguel P, O'Brien RM: **Multiple functional polymorphisms in the G6PC2 gene contribute to the association with higher fasting plasma glucose levels**. *Diabetologia* 2013, **56**(6):1306-1316.
225. Lieberman SM, Evans AM, Han B, Takaki T, Vinnitskaya Y, Caldwell JA, Serreze DV, Shabanowitz J, Hunt DF, Nathenson SG, Santamaria P, DiLorenzo TP: **Identification of the beta cell antigen targeted by a prevalent population of pathogenic CD8+ T cells in autoimmune diabetes**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8384-8388.
226. Strautnieks SS, Bull LN, Knisely AS, Kocoshis SA, Dahl N, Arnell H, Sokal E, Dahan K, Childs S, Ling V, Tanner MS, Kagalwalla AF, Nemeth A, Pawlowska J, Baker A, Mieli-Vergani G, Freimer NB, Gardiner RM, Thompson RJ: **A gene encoding a liver-specific ABC transporter is mutated in progressive familial intrahepatic cholestasis**. *Nat Genet* 1998, **20**(3):233-238.
227. Wang R, Salem M, Yousef IM, Tuchweber B, Lam P, Childs SJ, Helgason CD, Ackerley C, Phillips MJ, Ling V: **Targeted inactivation of sister of P-glycoprotein gene (spgp) in mice results in nonprogressive but persistent intrahepatic cholestasis**. *Proc Natl Acad Sci U S A* 2001, **98**(4):2011-2016.
228. Weil D, El-Amraoui A, Masmoudi S, Mustapha M, Kikkawa Y, Laine S, Delmaghani S, Adato A, Nadifi S, Zina ZB, Hamel C, Gal A, Ayadi H, Yonekawa H, Petit C: **Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin**. *Hum Mol Genet* 2003, **12**(5):463-471.
229. Ouyang XM, Yan D, Du LL, Hejtmancik JF, Jacobson SG, Nance WE, Li AR, Angeli S, Kaiser M, Newton V, Brown SD, Balkany T, Liu XZ: **Characterization of Usher syndrome type I gene mutations in an Usher syndrome patient population**. *Hum Genet* 2005, **116**(4):292-299.
230. Kitamura K, Kakoi H, Yoshikawa Y, Ochikubo F: **Ultrastructural findings in the inner ear of**

Jackson shaker mice. *Acta Otolaryngol* 1992, **112**(4):622-627.

231. Kikkawa Y, Shitara H, Wakana S, Kohara Y, Takada T, Okamoto M, Taya C, Kamiya K, Yoshikawa Y, Tokano H, Kitamura K, Shimizu K, Wakabayashi Y, Shiroishi T, Kominami R, Yonekawa H: **Mutations in a new scaffold protein Sans cause deafness in Jackson shaker mice**. *Hum Mol Genet* 2003, **12**(5):453-461.
232. Lukacs V, Mathur J, Mao R, Bayrak-Toydemir P, Procter M, Cahalan SM, Kim HJ, Bandell M, Longo N, Day RW, Stevenson DA, Patapoutian A, Krock BL: **Impaired PIEZO1 function in patients with a novel autosomal recessive congenital lymphatic dysplasia**. *Nat Commun* 2015, **6**:8329.
233. Fotiou E, Martin-Almedina S, Simpson MA, Lin S, Gordon K, Brice G, Atton G, Jeffery I, Rees DC, Mignot C, Vogt J, Homfray T, Snyder MP, Rockson SG, Jeffery S, Mortimer PS, Mansour S, Ostergaard P: **Novel mutations in PIEZO1 cause an autosomal recessive generalized lymphatic dysplasia with non-immune hydrops fetalis**. *Nat Commun* 2015, **6**:8085.
234. Zarychanski R, Schulz VP, Houston BL, Maksimova Y, Houston DS, Smith B, Rinehart J, Gallagher PG: **Mutations in the mechanotransduction protein PIEZO1 are associated with hereditary xerocytosis**. *Blood* 2012, **120**(9):1908-1915.
235. Albuisson J, Murthy SE, Bandell M, Coste B, Louis-Dit-Picard H, Mathur J, Feneant-Thibault M, Tertian G, de Jaureguiberry JP, Syfuss PY, Cahalan S, Garcon L, Toutain F, Simon Rohrlich P, Delaunay J, Picard V, Jeunemaitre X, Patapoutian A: **Dehydrated hereditary stomatocytosis linked to gain-of-function mutations in mechanically activated PIEZO1 ion channels**. *Nat Commun* 2013, **4**:1884.
236. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser**. *Genome Res* 2009, **19**(9):1630-1638.
237. Pemberton TJ, Wang C, Li JZ, Rosenberg NA: **Inference of unexpected genetic relatedness among individuals in HapMap Phase III**. *Am J Hum Genet* 2010, **87**(4):457-464.
238. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW: **Clines, clusters, and the effect of study design on the inference of human population structure**. *PLoS Genet* 2005, **1**(6):e70.
239. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, Weber JL: **Comparison of human genetic and sequence-based physical maps**. *Nature* 2001, **409**(6822):951-953.
240. Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ: **Sequence-level population simulations over large genomic regions**. *Genetics* 2007, **177**(3):1725-1731.
241. Fisher RA: **Statistical Methods For Research Workers**, 13th edn. New York: Hafner; 1925.
242. Matise TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, Hyland FCL, Kennedy GC, Kong X, Murray SS, Ziegle JS, Stewart WC, Buyske S: **A second-generation combined linkage-physical map of the human genome**. *Genome Res* 2007, **17**(12):1783-1786.
243. Fraley C, Raftery AE: **mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation.** In. Seattle, WA: Department of Statistics, University of Washington; 2012.
244. R Development Core Team: **R: A language and environment for statistical computing.** In. Vienna, Austria: R Foundation for Statistical Computing; 2017.
245. Hudson RR: **Generating samples under a Wright-Fisher neutral model of genetic variation**. *Bioinformatics* 2002, **18**(2):337-338.
246. Bhatia G, Patterson N, Sankararaman S, Price AL: **Estimating and interpreting FST: the impact of rare variants**. *Genome Res* 2013, **23**(9):1514-1521.
247. Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA: **Comparing spatial maps of human population-genetic variation using Procrustes analysis**. *Stat Appl Genet Mol Biol* 2010, **9**(1):Article 13.
248. Szpiech ZA, Hernandez RD: **selscan: an efficient multithreaded program to perform EHH-**

based scans for positive selection. *Mol Biol Evol* 2014, **31**(10):2824-2827.

249. Hintze JL, Nelson RD: **Violin plots: A box plot-density trace synergism.** *Am Stat* 1998, **52**(2):181-184.
250. Scott-Emuakpor AB: **The mutation load in an African population. I. An analysis of consanguineous marriages in Nigeria**. *Am J Hum Genet* 1974, **26**(6):674-682.
251. Coleman DA: **A note on the frequency of consanguineous marriages in Reading, England in 1972/1973**. *Hum Hered* 1980, **30**(5):278-285.
252. Freire-Maia N: **Inbreeding levels in American and Canadian populations: a comparison with Latin America**. *Eugen Q* 1968, **15**(1):22-33.
253. Jorde LB, Pitkanen KJ: **Inbreeding in Finland**. *Am J Phys Anthropol* 1991, **84**(2):127-139.
254. Stevenson AC, Johnston HA, Stewart MI, Golding DR: **Congenital malformations. A report of a study of series of consecutive births in 24 centres**. *Bull World Health Organ* 1966, **34 Suppl**:9-127.
255. Valls A: **Consanguineous marriages in a Spanish population**. *Acta Genet Stat Med* 1967, **17**(1):112-119.
256. Pinto-Cisternas J, Zei G, Moroni A: **Consanguinity in Spain, 1911-1943: general methodology, behavior of demographic variables, and regional differences**. *Soc Biol* 1979, **26**(1):55-71.
257. Freire-Maia N: **Inbreeding levels in different countries**. *Eugen Q* 1957, **4**:127-138.
258. Gomez PG: **Consanguinity: Geographical variation and temporal evolution in the North of the Iberian peninsula, 1918–1968 (León, Spain)**. *Int J Anthropol* 1989, **4**(1-2):119-124.
259. Du RF: **Percentages and types of consanguineous marriage in different nationalities of China**. *Zhonghua Yi Xue Za Zhi* 1981, **61**(12):723-728.
260. Zhang JX: **Effects of consanguineous marriages on hereditary diseases: a study of the Han ethnic group in different geographic districts of Zhejiang Province**. *Zhonghua Yi Xue Za Zhi* 1992, **72**(11):674-676, 703.
261. Liascovich R, Rittler M, Castilla EE: **Consanguinity in South America: demographic aspects**. *Hum Hered* 2001, **51**(1-2):27-34.
262. Orioli IM, Castilla EE, Carvalho WP: **Inbreeding in a South-American newborn series**. *Acta Anthropogenet* 1982, **6**(1):45-55.

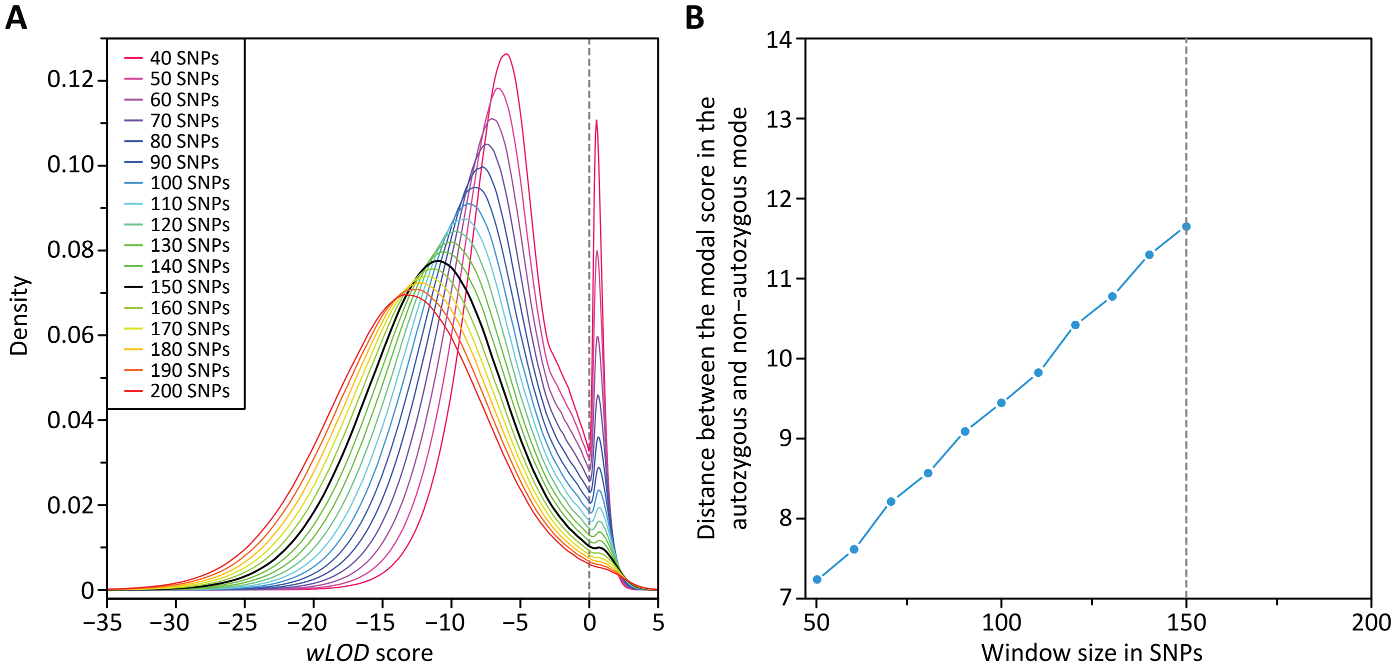

Figure 1. Distribution of genome-wide *wLOD* scores in European Americans. (**A**) Each line represents the Gaussian kernel density estimates of the pooled *wLOD* scores from all 97 individuals in the European American (CEU) population at window sizes between 40 and 200 SNPs in 10 SNP increments in the Omni2.5 dataset. The largest window size that produced a clear bimodal distribution (150 SNPs) is shown in black. (**B**) The change in intermodal distance with increasing window size in the CEU population. These patterns are representative of those observed in all other populations in the dataset.

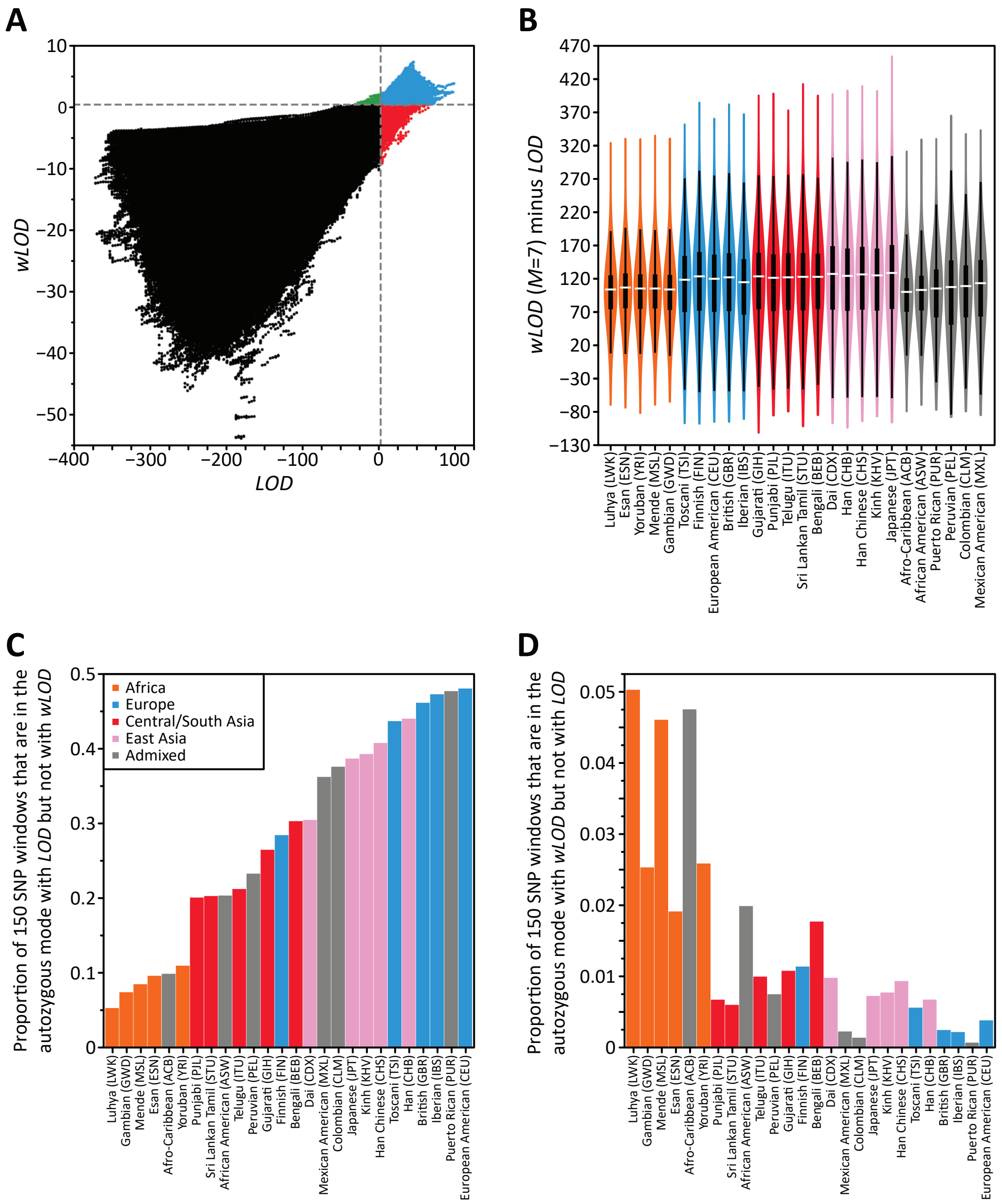

Figure 2. Difference in per-window scores between the LOD and $wLOD$ estimators. (**A**) Scatterplot comparing per-window LOD and $wLOD$ scores across all individuals in the European American (CEU) population at a window size of 150 SNPs in the Omni2.5 dataset. The 2,675,059 windows that were in the autozygous mode with both LOD and $wLOD$ are shown in blue. The 9,885 windows that were in the non-autozygous mode with LOD but the autozygous mode with $wLOD$ are shown in green. The 2,462,843 windows that were in the autozygous mode with LOD but the non-autozygous mode with $wLOD$ are shown in red. All windows that were in the non-autozygous mode with both LOD and $wLOD$ are shown in black. (**B**) Violin plots representing the change in per-window score between $wLOD$ and LOD across all individuals in each population for 150 SNP windows in the Omni2.5 dataset. Each ''violin'' contains a vertical black line (25%–75% range) and a horizontal white line (median), with the width depicting a 90°-rotated kernel density trace and its reflection, both colored by the geographic affiliation of the population [249]. Bar plots showing for each population the proportion of 150 SNP windows that are (**C**) in the autozygous mode with the LOD estimator but are in the non-autozygous mode with the $wLOD$ estimator or (**D**) in the non-autozygous mode with the LOD estimator but are in the autozygous mode with the $wLOD$ estimator.

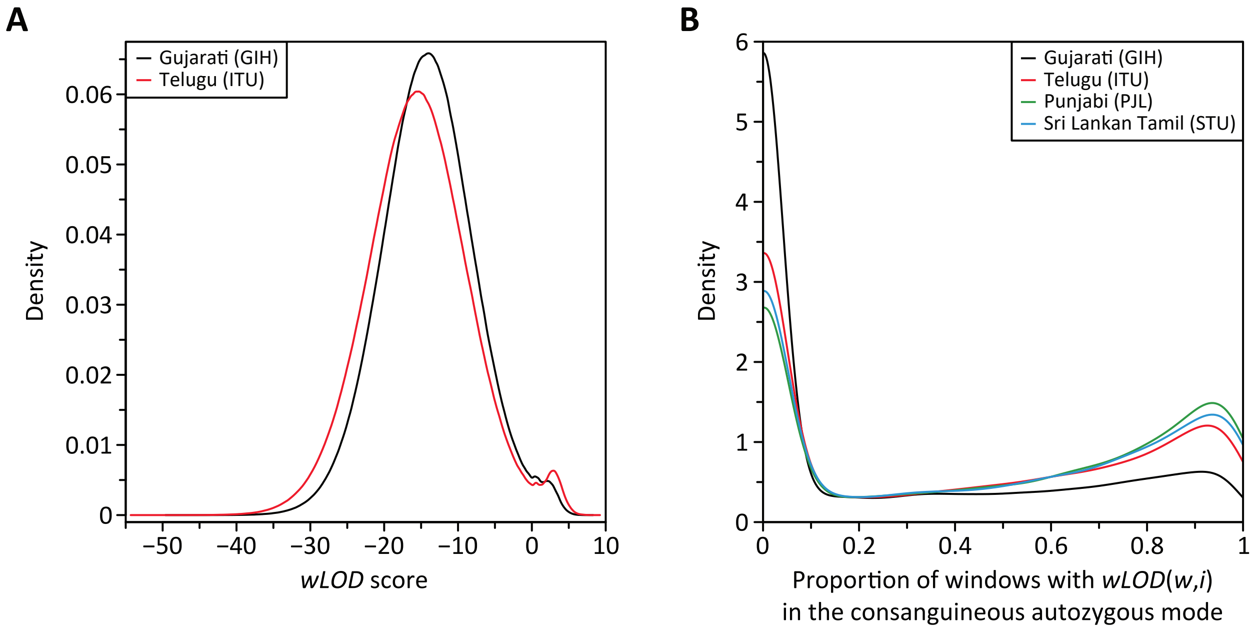

Figure 3. Influence of cultural processes on the distribution of *wLOD* scores. (**A**) Gaussian kernel density estimates of the pooled *wLOD* scores from all individuals in the Asian Indian Gujarati (GIH) and Telugu (ITU) populations at window sizes 200 and 220 SNPs, respectively. These patterns are representative of those observed in the Asian Indian Punjabi (PJL) and Sri Lankan Tamil (STU) populations as well as the East Asian Dai (CDX) population. (**B**) Gaussian kernel density estimates of the proportion of windows comprising each inferred ROA that are present in the right-most autozygosity mode in the Asian Indian GIH, ITU, PJL, and STU populations. ROA in the CDX population are almost exclusively in the left-most mode and it was excluded for clarity. The Asian Indian Bengali (BEB) population was excluded as we could not robustly distinguish between the two autozygous modes.

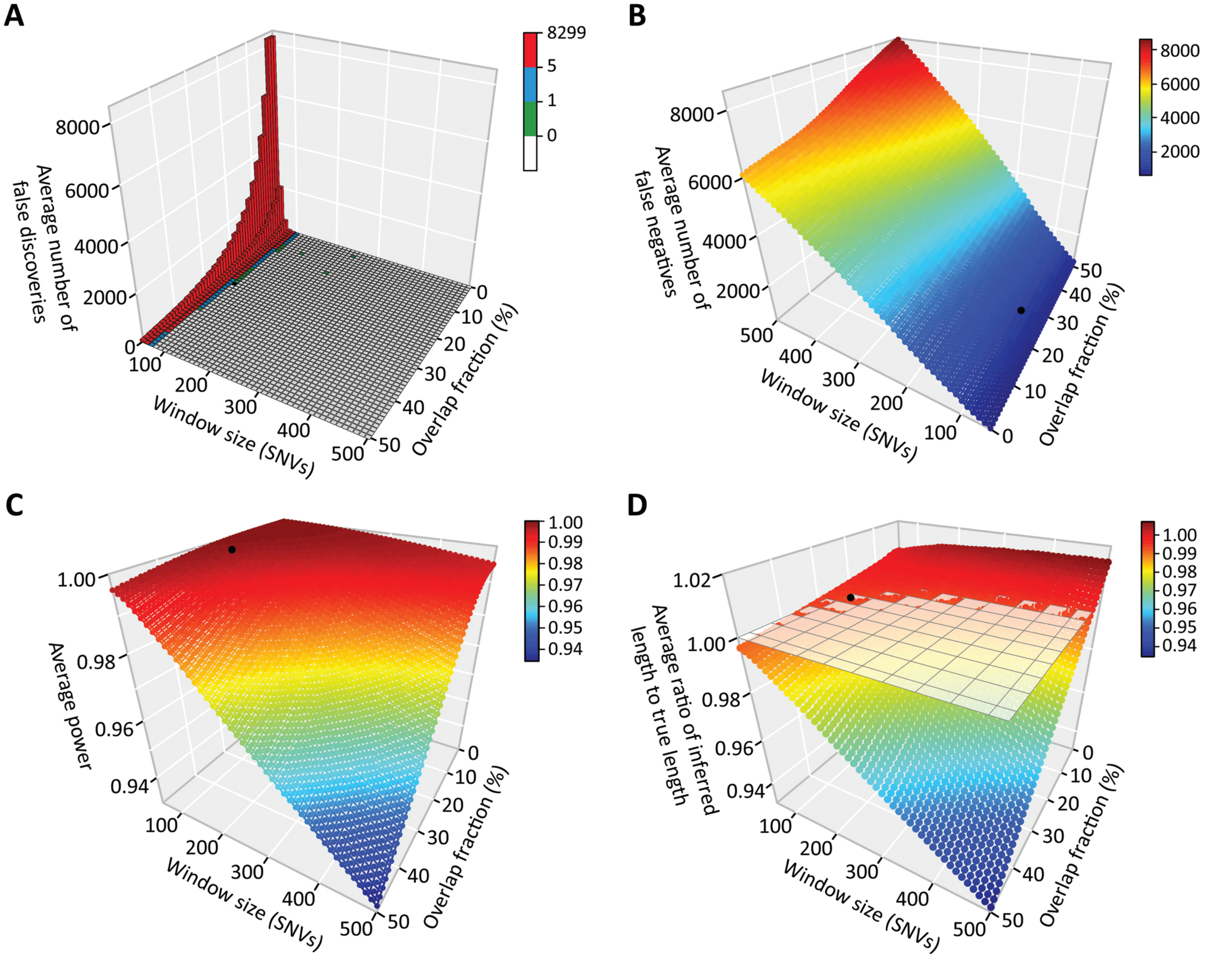

Figure 4. Performance of the *wLOD* method across different window sizes and overlap fractions. For scenario 1 and the 750,000 polymorphic SNV datasets, a three dimensional (3D) bar graph depicting the average number of falsely discovered ROA (**A**) as well as 3D scatterplots depicting the average number of false negative ROA (**B**), average power (**C**), and average ratio of inferred and true ROA lengths (**D**) reported by the *wLOD* method for each window size and overlap fraction across the 50 replicates are shown. In each graph, the point representing the smallest combination of window size and overlap fraction that had an average number of falsely discovered ROA of 0 and an average ratio of inferred and true ROA lengths of about 1 is shown in black.

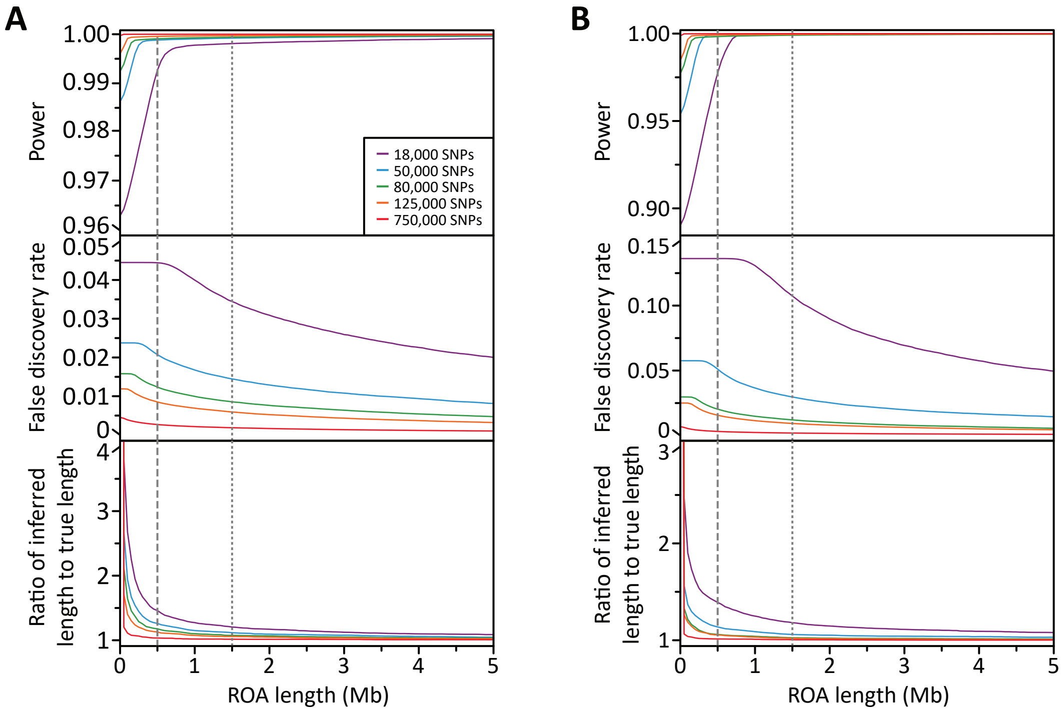

Figure 5. Performance of the *wLOD* method at different SNV densities. Lines graphs showing how average power (top), false discovery rate (middle), and ratio of inferred and true ROA length (bottom) across 50 replicate genetic simulations change with increasing ROA length for each SNV subset under (**A**) scenario 1 and (**B**) scenario 2. Each comparison was performed at the optimal combination of window size and overlap fraction for that scenario and SNV subset (**Table 3**). The grey vertical lines denote 500 kb (dashed) and 1.5 Mb (dotted), frequently applied length thresholds used to categorize ROA arising due to LD (< 500kb) and inbreeding (> 1.5 Mb) in humans [9]. Note that in scenario1, power to detect ROA >1 Mb with 18,000 SNVs surpasses that with 50-125,000 SNVs as a consequence of the optimal overlap fraction used: the overlap fraction of 0 used for the 18,000 SNV dataset is much lower than the 0.15–0.22 fractions used for the 50-125000 SNV datasets. Consequently, greater power to detect ROA >1Mb is achieved with 18,000 SNVs than is possible with 50-125,000 SNVs through less stringent placement of ROA boundaries, but at the expense of more frequent overcalling of ROA (inflated false discovery rate).

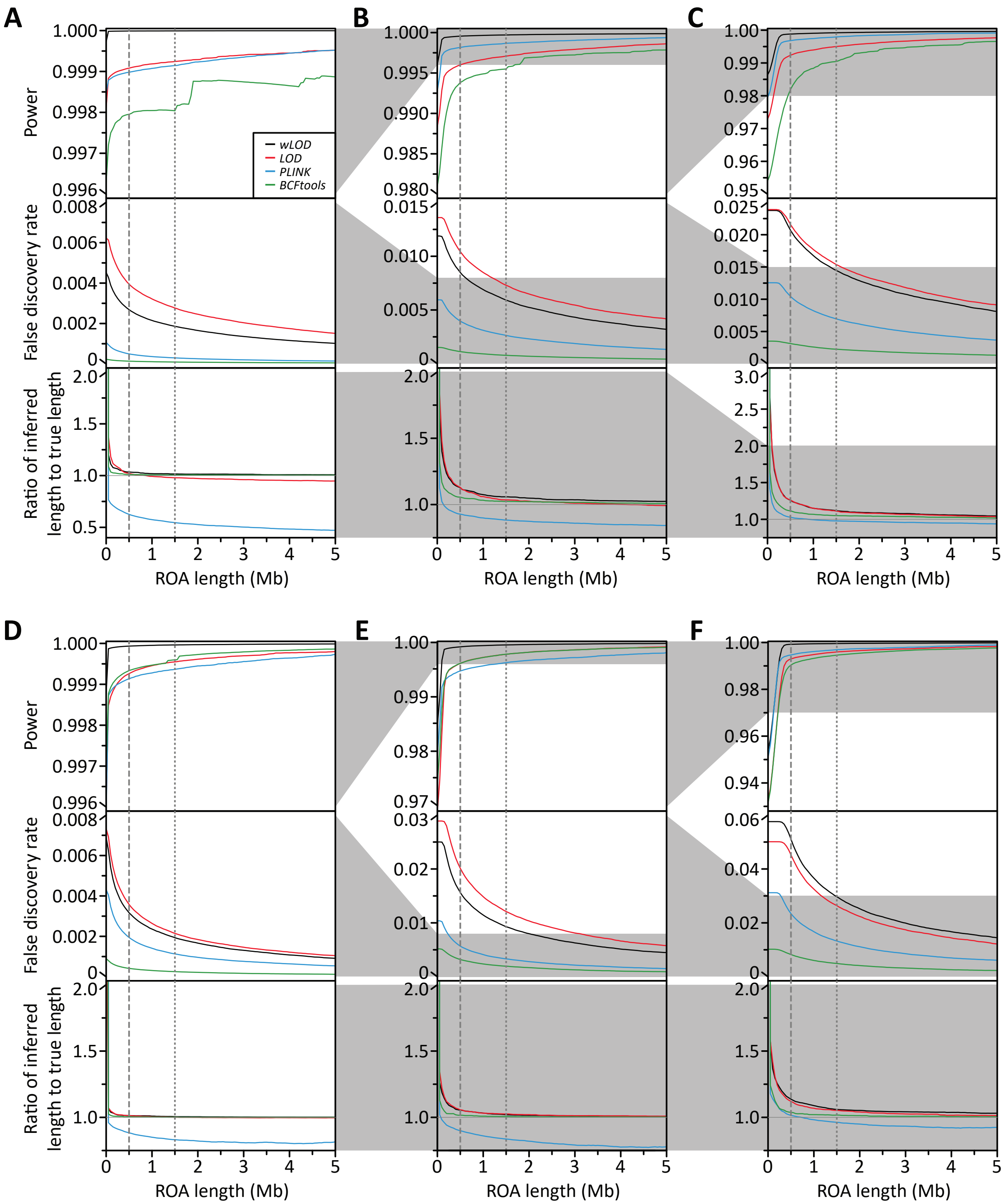

Figure 6. Performance of the *wLOD* method compared with existing methods. Line graphs showing for scenarios 1 (**A-C**) and 2 (**D-F**) and subsets consistent with WGS (750,000 SNV; **A & D**) and the Illumina HumanOmni2.5-8 (125,000 SNV; **B & E**) and HumanOmniExpress-24 (50,000 SNV; **C & F**) BeadChips how average power (top), false discovery rate (middle), and ratio of inferred and true ROA length (bottom) across 50 replicate genetic simulations change with increasing ROA length. The grey vertical lines denote 500 kb (dashed) and 1.5 Mb (dotted), frequently applied length thresholds used to categorize ROA arising due to LD (< 500kb) and inbreeding (> 1.5 Mb) in humans [9].

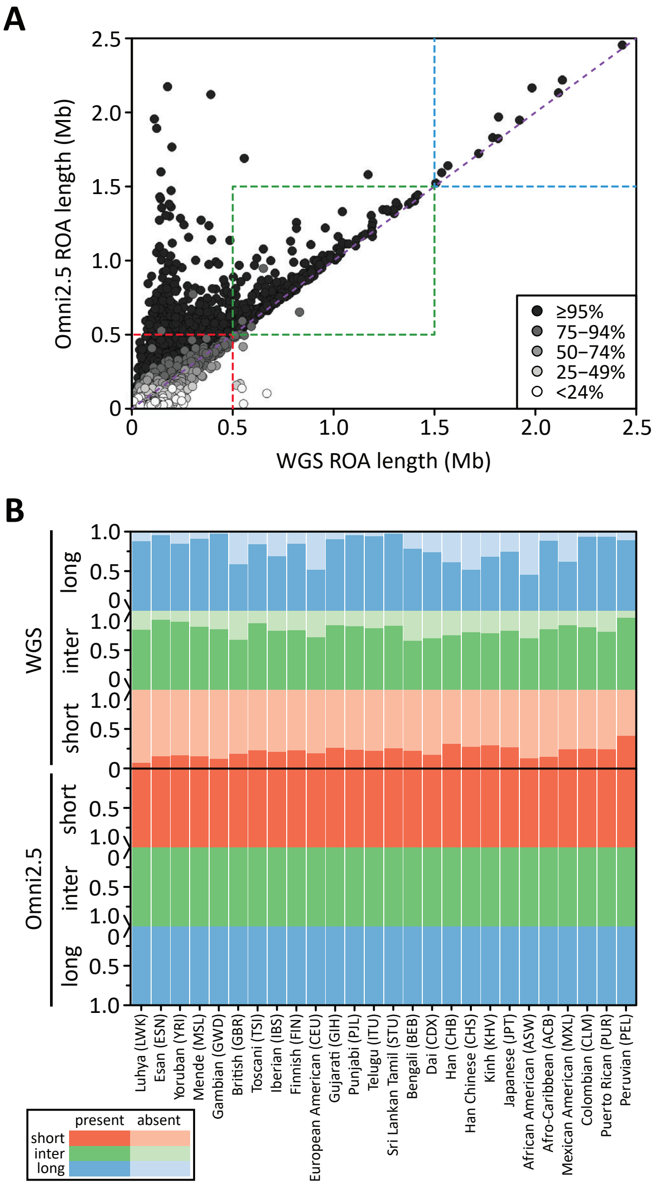

Figure 7. Concordance of ROA inferred in the WGS and Omni2.5 datasets. (**A**) A scatterplot comparing the length of each WGS ROA with that of its corresponding Omni2.5 ROA in the European American (CEU) population. Each point is shaded according to the proportion of the WGS ROA that overlaps the Omni2.5 ROA. (**B**) Bar plots representing the proportions of short (<500 kb; shown in red), intermediate (500 kb to 1.5 Mb; shown in green), and long (>1.5 Mb; shown in blue) ROA in the WGS (upper) and Omni2.5 (lower) datasets that overlap (darkest shade) or are absent from (lightest shade) the other dataset in each population.

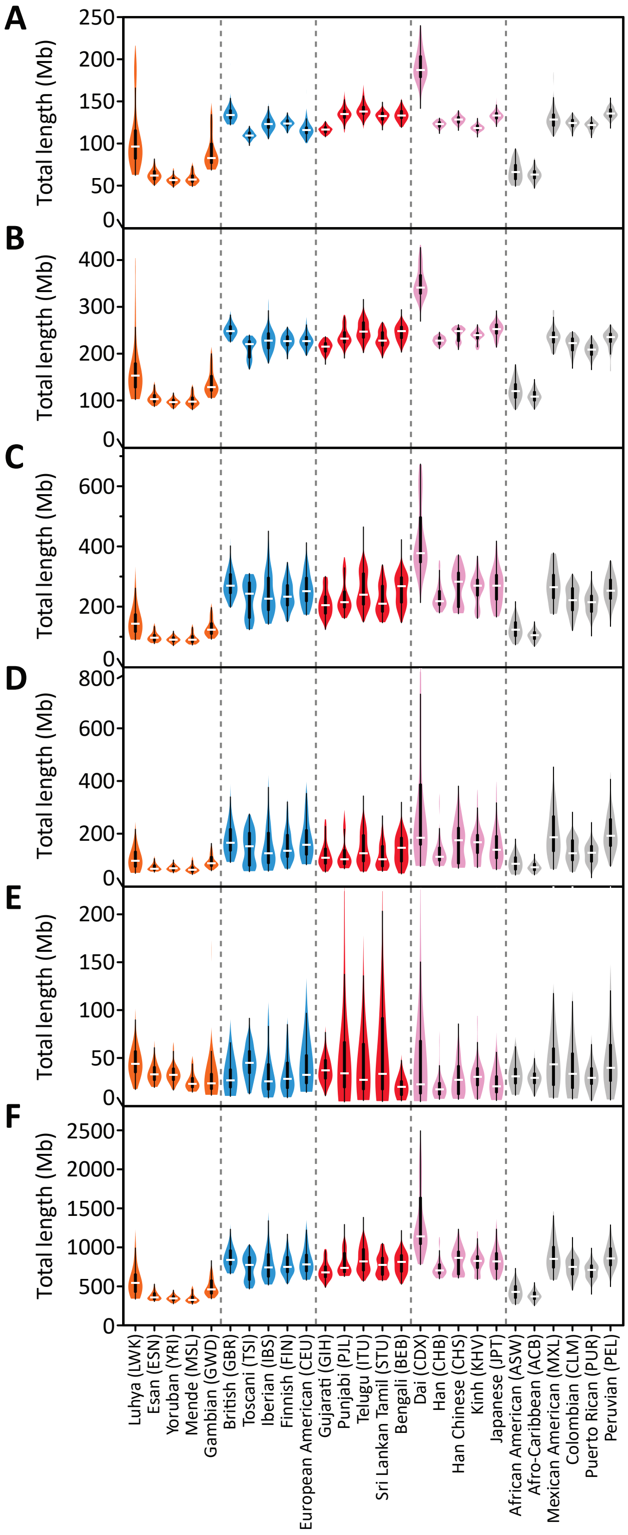

Figure 8. Population-specific distributions of the total length of ROA per individual. Data are shown as violin plots [249], representing the distribution of total ROA length across all individuals in each of the 26 populations for (**A**) class 1, (**B**) class 2, (**C**) class 3, (**D**) class 4, (**E**) class 5, and (**F**) all five ROA classes combined. Populations are ordered from left to right by geographic region and within each region by increasing geographic distance from Addis Ababa.

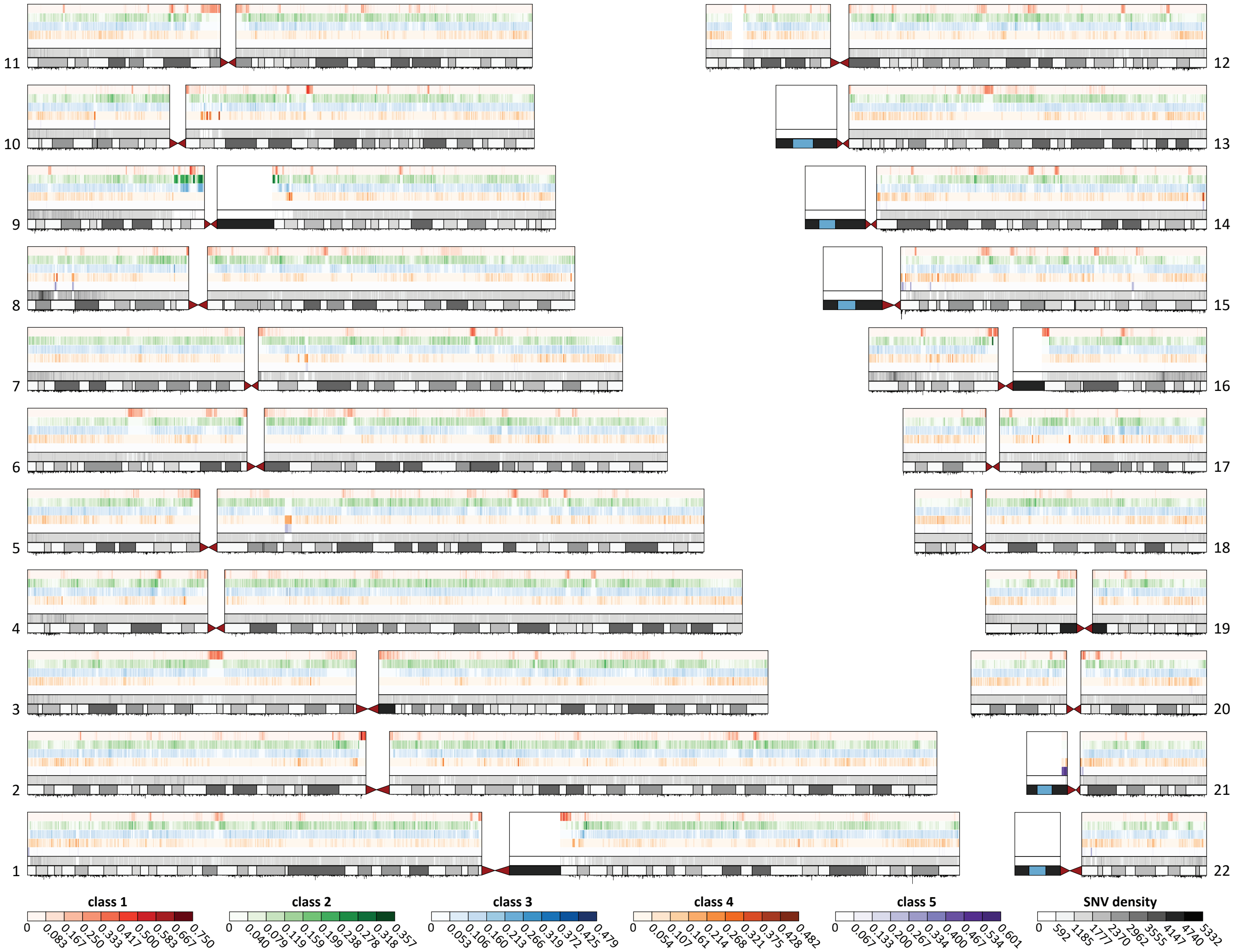

Figure 9. Distribution of worldwide ROA frequencies across the genome. For each autosome, the figure shows for each ROA length class the average proportion of individuals in the WGS dataset who have an ROA overlapping SNVs within non-overlapping 50 kb windows. Each row represents an ROA class, and each column represents a window. The intensity of a point increases with increasing average ROA frequency, as indicated by the color scale below the figure. The SNV density of each window and an ideogram of chromosome banding are shown in the bottom tracks, with average recombination rate in each window represented by a vertical black line below the ideogram, where line heights proportional to average recombination rate.

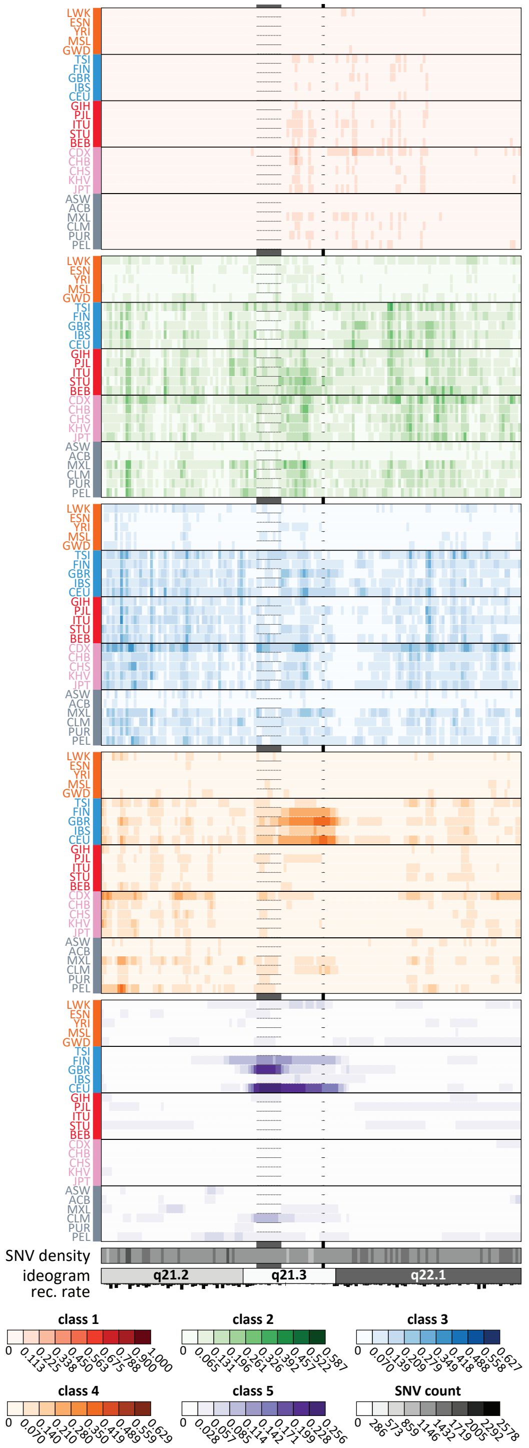

Figure 10. Per-population ROA frequencies within a ROA hotspot on chromosome 2. For each ROA class, for each population, the average proportion of individuals in that population who have an ROA overlapping SNVs within non-overlapping 50 kb windows from 132,500,000 to 140,200,000 bp on the q-arm of chromosome 2 is shown. Each row represents a population, and each column represents a window. Populations are ordered from top to bottom by geographic affiliation, as indicated by the color of their label, and within regions from top to bottom by increasing geographic distance from Addis Ababa (in the same order as in **Figure 8**). Average ROA frequency, average SNV density, chromosome banding, and recombination rates are shown as in **Figure 9**. The black vertical box demarks the location of the *LCT* gene, while the vertical grey box demarks the location of the class 5 ROA hotspot in the CEU and GBR.

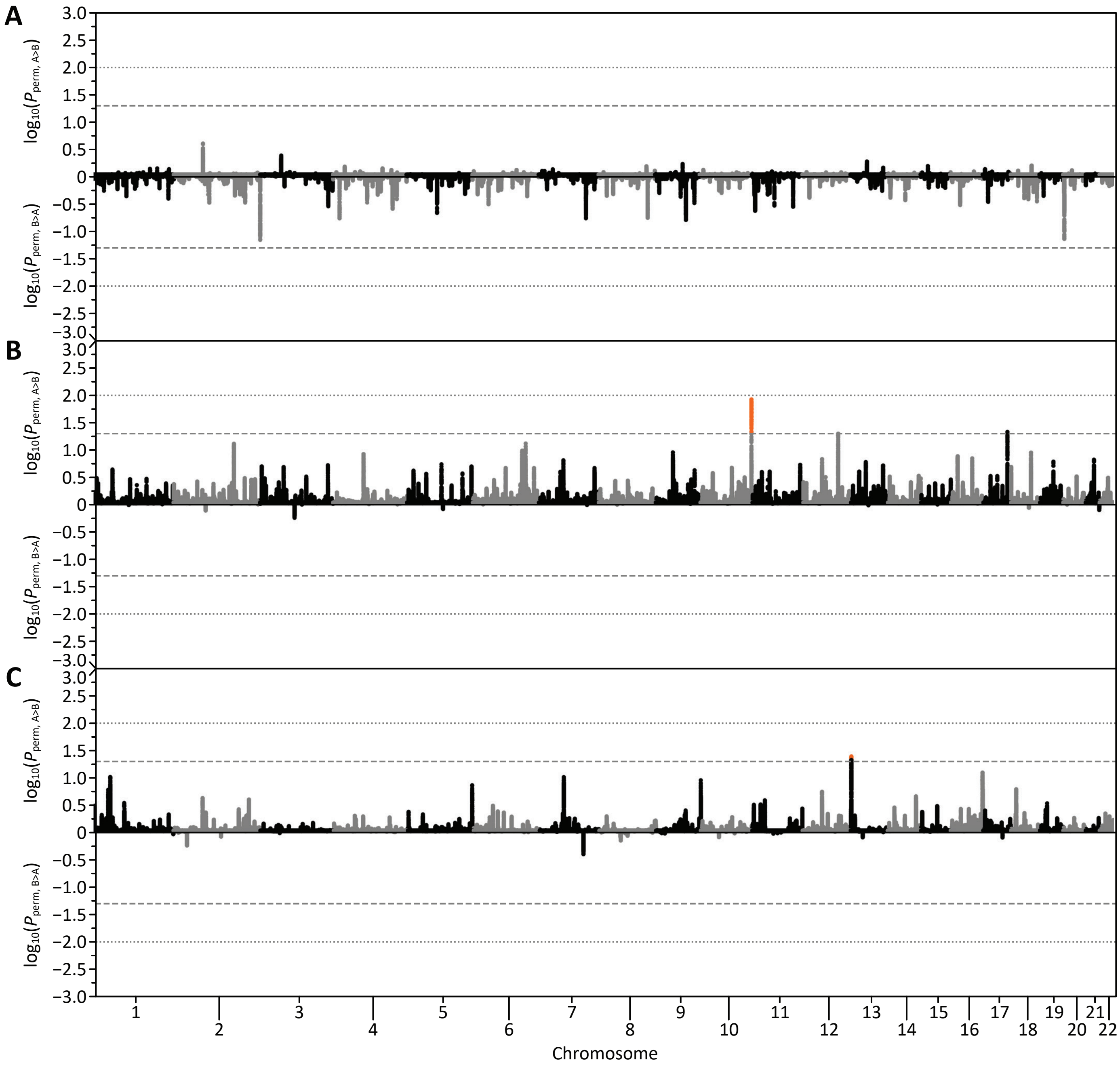

Figure 11. Distribution of differential ROA signals between subgroups in the GIH, ITU and PJL. Manhattan plots showing for each window the $log_{10}(P)$ of pairwise comparisons of per-individual $wLOD$ scores in the two subgroups present in the (**A**) GIH (450 SNV window), (**B**) ITU (580 SNV window), and (**C**) PJL (610 SNV window). In each plot, *P*-values for the comparison testing whether $wLOD$ scores in cluster A are greater than those in cluster B (Additional File 1: **Figure S22**) are shown on top with *P*-values for the reverse comparison is shown below. *P*-values represent the proportion of genome-wide maximum Wilcoxon rank-sum test statistics observed in 1,000 permutations of group labels that exceed the Wilcoxon rank-sum test statistic obtained with the true labels [150]. Windows with *P*>0.05 are shown in black and those with *P*≤0.05 are shown in orange. The horizontal grey dashed line denotes *P*=0.05 and while the grey dotted line denotes *P*=0.01.

Table 1. Per-SNP likelihoods of autozygosity and non-autozygosity.

$\boldsymbol{G_k}$	$\mathbf{Pr}(\boldsymbol{G_k} \mid \boldsymbol{X_k} = \mathbf{1})$	$\mathbf{Pr}(\boldsymbol{G_k} \mid \boldsymbol{X_k} = \mathbf{0})$
AA	$(1-\varepsilon)f_{A,j} + \varepsilon f_{A,j}$	$f_{A,j}^2$
AB	$2\varepsilon f_{A,j} f_{B,j}$	$2 f_{A,j} f_{B,j}$
BB	$(1-\varepsilon)f_{B,j} + \varepsilon f_{B,j}$	$f_{B,j}^2$
Missing	1	1

Frequencies of alleles A and B in population j are denoted by $f_{A,j}$ and $f_{B,j}$, respectively, and the assumed rate of genotyping errors and mutations by ε.

Table 2. Populations included in Phase 3 of The 1000 Genomes Project.

Population		Geographic region	*N*	Consanguinity[a]	
ID	Name			Frequency	Reference(s)
ESN	Esan	Africa	94	–	–
GWD	Gambian	Africa	109	–	–
LWK	Luhya	Africa	96	–	–
MSL	Mende	Africa	80	–	–
YRI	Yoruban	Africa	107	51.20%	[250]
GBR	British	Europe	89	0.40%	[251]
CEU	European American	Europe	97	0.20%	[252]
FIN	Finnish	Europe	98	0.17%	[253]
IBS	Iberian	Europe	107	1.99%	[254-258]
TSI	Toscani	Europe	104	–	–
BEB	Bengali	Central/South Asia	84	5.00%	[173]
GIH	Gujarati	Central/South Asia	101	4.90%	[173]
PJL	Punjabi	Central/South Asia	96	0.90%	[173]
STU	Sri Lankan Tamil	Central/South Asia	96	38.20%	[173]
ITU	Telugu	Central/South Asia	101	30.80%	[173]
CDX	Dai	East Asia	92	21.30%	[172]
JPT	Japanese	East Asia	103	4.80%	
KHV	Kinh	East Asia	94	–	–
CHB	Northern Han	East Asia	101	1.16%	[172,259,260]
CHS	Southern Han	East Asia	102	3.43%	[172,260]
ASW	African American	Admixed	55	–	–
ACB	Afro-Caribbean	Admixed	94	–	–
CLM	Colombian	Admixed	89	2.83%	[252,254,261]
MXL	Mexican American	Admixed	62	0.80%	[252,254]
PEL	Peruvian	Admixed	84	1.90%	[252,261,262]
PUR	Puerto Rican	Admixed	101	3.30%	[257]

[a]Consangunity frequencies were obtained from http://www.consang.net.

Table 3. Optimal window size and overlap proportion in the simulated datasets.

SNV subset	Scenario 1		Scenario 2	
	Window size	**% overlap**	**Window size**	**% overlap**
18,000	60	7	70	5
50,000	70	18	80	19
80,000	80	28	90	22
125,000	80	29	100	26
750,0000	130	37	120	32

Table 4. Genomic regions enriched for autozygosity signals in the ITU and PJL subgroups.

Population			**Genomic region**				**Number of windows**	**Minimum P_{perm}**	**RefSeq genes**
ID	**Name**	**Group**	**Chr**	**Begin (bp)**	**End (bp)**	**Length (bp)**			
ITU	Telugu	1	10	132,953,074	133,048,305	95,232	189	0.013	*TCERG1L*
PJL	Punjabi	1	13	20,001,572	20,181,691	180,120	28	0.044	*TPTE2*