The Effect of Ascertainment Bias on Detecting Signatures of Selection

Marla Willemse

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A Dissertation submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine.

Johannesburg, 2019

Declaration

I, Marla Willemse, declare that this thesis is my own work. It is being submitted for the degree of Master of Science (Medicine) at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree at any university.

_____

Signature

_____

Date

Abstract

Genotyping arrays have been broadly used to identify signatures of selection with genome-wide scans. It has been reported that the markers contained in arrays don't accurately represent the variation in full sequence data, especially in non-European populations, and that this may affect the results of selection studies. The availability of whole genome sequence (WGS) data from various African populations has enabled the analysis of the extent to which ascertainment bias affects the detection of selection signals on this continent.

Seven commonly used genotyping arrays were represented by creating *in silico* single nucleotide polymorphism (SNP) panels from WGS data of the African Genome Variation Project (AGVP) Baganda, Ethiopia and Zulu samples. Four types of selection scans (F_{ST}, iHS, XP-EHH and Tajima's D) were performed on both the array and WGS datasets, and the accuracy of selection signals identified from array data was assessed in relation to the WGS results.

It was found that selection scans performed with array data produced a significant proportion of false positives and false negative signals. The EHH-based methods were least affected by ascertainment bias and arrays with higher marker density generally produced more accurate results. The two arrays ascertained from African populations out-performed a more European-based array of similar size.

Variation in marker density across the genome was found to underlie discrepancies between array and WGS selection signals, as genomic regions in array data containing fewer markers were less likely to be detected as selection signals. Of the selection signals identified from WGS but not array data, most were missed due to insufficient SNP density.

To investigate the extent to which the selection signals from one Southeastern Bantu-speaking (SEB) group is shared by another SEB group, selection scans on two independent SEB groups, namely the Bt20 and AGVP Zulu samples. The overlap in

selection signals between the samples was found to be limited, concurring with differential KhoeSan gene flow into these groups.

It was found that various selection scan methods are differentially affected by ascertainment bias, and additionally, limited concordance was observed between the selection signals identified by different methods. A comparison of selection signals between the three AGVP populations revealed high population specificity of signals.

Regions displaying signatures of selection were annotated for gene names and functionality, and both canonical and less well-established selection candidates were identified. These included genes associated with infectious diseases, cancer, metabolism, pigmentation, neuro-motor functions and high altitude adaptation.

Acknowledgements

This project would not have been possible without my supervisors, Zané Lombard, Dhriti Sengupta and Shaun Aron. Thank you for your invaluable and patient guidance. I also owe great thanks to Ananyo Choudhury for sharing his vast knowledge freely.

I extend my appreciation to the SBIMB for investing in its student and to my colleagues for being genuine and caring people.

I would like to thank my friends and family for making me consolatory popcorn when my scripts wouldn’t run, and the kind strangers on Stack Overflow for fixing the deviant code.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Finally, I would like to thank the universe for being weird and interesting.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Affy 6:	Affymetrix 6.0 array
AGVP:	African Genome Variation Project
Bt20:	Birth to Twenty
EHH:	extended haplotype homozygosity
FN:	false negative
FP:	false positive
FPR:	false positive rate
F_{ST}:	Wright's fixation index
H3A:	Illumina H3A array
iHS:	integrated haplotype score
kb:	kilobase
KGP:	1000 Genomes Project
LD:	linkage disequilibrium
M:	millions of SNPs
MAF:	minor allele frequency
Omni 1:	Illumina Omni 1.0 array
Omni1 2.5:	Illumina Omni 2.5 array
Omni 5:	Illumina Omni 5.0 array
PanAFR:	Affymetrix PanAFR

R factor:	representation factor
SEB:	Southeastern Bantu-speaking
SFS:	site frequency spectrum
SNP:	single nucleotide polymorphism
TN:	true negative
TNR:	true negative rate
TP:	true positive
TPR:	true positive rate
XP-EHH:	cross-population extended haplotype homozygosity
WGS:	whole genome sequence

1. Introduction

1.1. Molecular evolution and selection

The genomic variation of a population changes from one generation to the next, resulting in phenotypic transitions over a large timescale. This can take place due to neutral effects such as genetic drift, in which the random sampling of the parental generation's genetic variation contributes to the successive generation. Other neutral evolutionary influences include demographic effects such as population size changes and migrations, which can introduce or remove variation (Bazin et al. 2010). Adaptive evolution occurs when genetic variants differentially affect survival in a certain environment, causing changes in a population's genomic composition. Most mutations in functional genomic regions are mildly deleterious, negatively affecting survival, and are removed by purifying selection (Bank et al. 2014). Conversely, positive directional selection occurs when a genetic variant confers a selective advantage, causing it to propagate more rapidly (Orr 2009). As the selected variant 'sweeps' to high frequency, a detectable pattern of allelic variation is created in the surrounding region: a signature of selection (Chen et al. 2010). The availability of whole genome sequence (WGS) and high density genotyping array data has enabled genome-wide scans for signatures of selection. These patterns are studied to form hypotheses of evolutionary histories and identify functional genomic regions.

1.2. Histories of populations studied

Selection in Africa is interesting both because of the length of human history on the continent and the diversity of its ethnic groups. A debate rages around the origin of modern humans, as evidence is continually uncovered. The 'Single African Origin' theory holds that modern humans originated in Africa around 200 thousand years ago (kya) before migrating within and out of Africa to populate the world (Campbell & Tishkoff 2010). The contesting 'African Multiregional' model draws on evidence that modern human traits evolved in multiple locations within Africa (Henn et al. 2018).

The Out of Africa migration 50-100kya involved a bottleneck which reduced genomic variation in non-African populations (Campbell et al. 2014). The higher levels of diversity both within and between African populations are consistent with the large population sizes and substructure maintained throughout their long history on the continent. It should be noted that 'population' is a loose term which can be defined by genetic, geographic or linguistic affiliation, at various levels of similarity.

1.2.1. Origins, migrations & languages

The genetic relatedness of populations is loosely correlated with both the similarity of their languages and their geographic proximity (Tishkoff et al. 2009; Campbell et al. 2014; Gomez et al. 2014). The diversity of Africa is reflected by its estimated 2,000 languages, which can be grouped into four families: KhoeSan, Nilo-Saharan, Niger-Kordofanian and Afroasiatic (Gomez et al. 2014).

The Niger-Kordofanian family consists of agro-pastoralists from sub-Saharan Africa. The Bantu-speaking branch of this superfamily includes the Zulu- and Luganda- speaking groups (Tishkoff et al. 2009; Campbell et al. 2014) which were examined in this study. One of the greatest demographic upheavals on the continent occurred 3-5kya when waves of Bantu-speaking agropastoralists from the the Grasfields region, located within the current-day borderlands of Nigeria and Cameroon migrated throughout sub-Saharan Africa, arriving in southern Africa ~1.5kya (Chimusa et al. 2015). Eastern Bantu-speaking populations were formed by two admixture events between the western Bantu speakers and an Afro-Asiatic speaking population from Ethiopia, occurring 1-1.5kya and 150-400 years ago (Patin et al. 2017). The Bantu expansion 2-3kya resulted in the spread of agriculture and admixture with pre-occurring populations (Beltrame et al. 2016). Today, a majority (~70%) of southern Africans belong to the Bantu language group. The Southeastern Bantu-speaking (SEB) group referred to in this study is a diverse collection of populations which have diverged and admixed since the Bantu expansions into Southeastern Africa. The Zulu population which is part of the SEB group derives about 23% of its ancestry from recent (<1kya) admixture with the KhoeSan, the earliest diverging human population (Gronau et al 2011; Veeramah et al. 2011 and Schlebusch et al. 2012).

The third group of populations represented in this study consists of the Oromo, Amharic, Somali and Wolaytta speaking people of Ethiopia. The Amharic language group belongs to larger Semitic group, the Oromo and Somali to the Cushitic group, and the Welayta to the Omotic group. These languages belong to the Afro-Asiatic super-group, which is spoken in northern and eastern Africa. Afro-Asiatic speakers originally ranged from the Nile Valley to the Ethiopian highlands, before migrating east, west and north 8-5kya (Tishkoff et al. 2009). Back-migrations from Eurasia 2.7- 3.3kya as well as admixture with the Khoe-San contributed to the genomic variation of eastern Africans (Pickrell et al. 2014). The divergence between the populations examined in this study is characterised by genome-wide average FST estimates, which were generated by the African Genome Variation Project (AGVP) (Gurdasani et al. 2015 supplemental information)

Table 1.1 F_{ST} Estimates from WGS Data Generated by the AGVP Study

Population	Baganda	Zulu	Oromo	Amhara	Somali
Baganda	0.000	0.008	0.035	0.039	0.035
Zulu	0.009	0.000	0.041	0.045	0.041
Oromo	0.025	0.032	0.000	0.000	0.008
Amhara	0.025	0.032	0.002	0.000	0.009
Somali	0.037	0.044	0.018	0.017	0.000

African genomes have been shaped by this complex demographic history, together with adaptation to a range of environments, climates, diets and pathogen exposures.

1.2.2. Previously identified signatures of selection in Africa

Hundreds of genomic regions displaying signatures of selection have been detected, providing insight into historic adaptations. Following the advent of agriculture 10-12kya in the Middle East, global human populations underwent growths and migrations (Atkinson et al. 2009). These changes drove the spread of infectious diseases, while exposing populations to a range of environments and selective pressures (Voight et al. 2006). These recent changes are reflected in African genomes by the enrichment of selected alleles which are at lower frequencies than in non-African genomes (Liu et al. 2013). Selection on morphological and reproductive features is known to act on a larger timescale, while selection on metabolic and immune-related genes occurs in response to recent environmental changes (Voight et al. 2006). The strongest selective pressures

are expected to produce the clearest genetic signals, and some of the most well-known signatures of selection include regions related to immunity and metabolism (Sègurel et al. 2017; Vatsiou et al. 2016a).

Infectious diseases are believed to have been one of the strongest drivers of recent human adaptation (Wills & Green 1995; Vatsiou et al. 2016a). Studies of population-specific variation and signatures of selection have revealed possible causes of differences in disease susceptibility between populations, and accounted for the high frequencies of some deleterious alleles (Daub et al. 2013). Selection on variants which contribute immunity to infectious diseases may have simultaneously caused noninfectious diseases to rise to a high frequency (Gomez et al. 2014). Chimusa et al. (2015) found that many selection candidates in Southern Africa were associated with infectious diseases such as influenza, tuberculosis, and HIV/AIDs and malaria.

A variety of subsistence strategies have been practiced in African history, including hunting and gathering, agriculture, and pastoralism. Dietary changes have presented strong selective pressures, and novel dietary adaptations include bitter taste perception, iodine metabolism, amylase copy number and lactase persistence (Beltrame et al. 2016). These adaptations can be reconstructed by examining the characteristic patterns of genomic variation which they create.

1.3. Signatures of selection

As a selective sweep rapidly increases the prevalence of a selected allele, recombination is unable to break up the association with surrounding variants at a corresponding pace, leading to hitch-hiking of genomic segments with the selected allele. This association between closely located alleles is called linkage disequilibrium (LD) and the collection of alleles which occur together due to their proximity constitutes a haplotype (Reich et al. 2001; Sabeti et al. 2002). During a selective sweep, neighbouring variants ‘hitchhike’ to high frequencies together with the selected allele. This creates a long region of depleted variation, called extended haplotype homozygosity (EHH) (Sabeti et al. 2002).

New mutations eventually restore diversity, but these are initially present at low frequencies. Most novel mutations in functional regions are disadvantageous and are quickly removed by negative selection (also called purifying or background selection), creating highly conserved sequences. However, when neutral or mildly deleterious mutations occur in a region strongly affected by positive selection, they are 'hidden' from negative selection (Lewontin & Krauker 1973). This continues until the positively selected allele becomes fixed, occurring throughout the population and no longer having a relative advantage (Zhai et al. 2009). Thus, an excess of rare variants appears on the selected haplotype. This causes a shift in the distribution of allele frequencies, called the site frequency spectrum (SFS), away from intermediate frequency alleles (Tajima 1989).

These effects are specific to the population experiencing a certain selective pressure. As populations in different environments experience distinct environmental influences, the divergence at selected sites becomes higher than at the genome-wide level. This increase in population differentiation can be detected as a signature of selection (Bank et al. 2014). Many methods to detect selection have been developed, all of which rely on certain characteristics of a sweep, detect selection at various time depths and have different assumptions. Creating robust methods to identify selection has proven challenging, and selection studies are packaged with a large index of limitations, confounders and caveats (Pavlidis et al. 2012; Fagny et al 2014; Vatsiou et al. 2016b).

1.3.1. Variations of positive selection

Selection does not always leave such a legible signature, and other modes of selection create complex patterns which are less detectable by genome scans (Enard et al. 2014). Balancing selection occurs when the heterozygote genotype is favoured. As both alleles are maintained in a population, the action of a selective pressure is less noticeable by comparison to frequencies of alternate alleles (Schrider & Kern 2017).

Signatures of selection can also be obscured in the case of a soft selective sweep. When selection favours a novel beneficial mutation, the single haplotype on which the new variant occurs is driven to a high frequency in a 'hard sweep'. However, an environmental change can cause a pre-existing variant to gain a selective advantage. Since this standing variation occurs on multiple haplotypes backgrounds, a soft sweep

produces a less clear signal, which most methods are underpowered to detect (Colonna et al. 2014; Vatsiou et al. 2016b). Schrider & Kern (2017) found that 92% of selective sweeps are soft, and that this proportion is significantly higher in African populations, as is expected since they have more standing variation for selection to act on. This means that selection in humans is not mutation-limited, and populations are able to adapt to changing conditions more rapidly, without needing to wait for a beneficial mutation to occur.

The stronger a selective pressure, the faster the sweep takes place, and the less time there is for recombination to break up the haplotype. Thus, a stronger selective pressure will produce a longer region of EHH, which is easier to detect. However, in a larger haplotype block it is more difficult to distinguish the causal single nucleotide polymorphism (SNP) from hitchhiking variants (Sabeti et al. 2006).

1.3.2. Detecting signatures of selection

Many methods have been developed to detect signatures of selection, and each assigns a value of a summary statistic to a genomic window. Regions which display the characteristics of selection are expected to produce the most extreme values of these statistics. There are two possible approaches to detect regions with extreme values. The first option is to simulate genomic data under the standard neutral population model, which characterizes genomic variation in the absence of selection, to serve as a null hypothesis (Zhai et al. 2009). Real test data is then compared to provide p values indicating the likelihood of selection. The validity of the simulation approach is limited by its reliance on correctly specifying historic population parameters, which are not well characterized (Nielsen et al. 2005; Pavlidis et al. 2012; Ferrer-Admetlla et al. 2014).

Alternatively, the outlier approach involves identifying selection candidates as the genomic regions with the highest values of a statistic, compared to the genome-wide empirical distribution (Manel et al. 2016). The genome is divided into windows of a predefined size, and a certain percentage of windows is chosen as an outlier threshold. Windows are ranked by their average value for the statistic and those with the most extreme scores are identified as outliers in the empirical distribution. The choice of window size is also not standard and variations between studies may affect results.

An advantage of the outlier approach is that results are more comparable between different methods (Ferrer-Admetlla et al. 2014). However, it's unknown what percentage of the genome is truly affected by selection, so outlier thresholds are chosen arbitrarily. The outlier approach isn't a formal significance test and will almost inevitably under- or over-estimate the percentage of the genome which is under selection (Granka et al. 2012; Pavlidis et al. 2012). The proportion of false positive signals is still unknown and may be large (Teshima et al. 2006; Pickrell et al. 2009). Since they occupy a proportion of the tail of the empirical distribution, the number of false positives reduces the number of true positives (Pavlidis et al. 2012).

A major challenge in identifying variants which are truly under positive selection is to distinguish the effects of this form of selection from other processes which produce similar changes in allele frequencies. Demographic effects which mimic the effect of selection include population expansions, bottlenecks, migrations, admixtures and subdivisions, and the proportional effect of various confounders is still unknown (Bank et al. 2014; Cadzow et al. 2014; Colonna et al. 2014). The outlier approach assumes that demographic effects produce uniform changes in variation throughout the genome, while selection produces local effects which stand out against genome-wide patterns (Manel et al. 2016). This assumption is generally true, but exceptions to the rule may result in many false classifications.

1.4. Methods to detect selection

Signatures of selection are characterized by an increase in population differentiation, extended haplotype homozygosity, and a shift in the SFS. Many selection statistics have been developed, and four were chosen for this study, each of which detects a different attribute of selective sweeps.

1.4.1. F_{ST}

Wright's fixation index (F_{ST}) detects geographically localized selection which causes high divergence at a given site relative to the distribution across the genome (Colonna et al. 2014). F_{ST} compares the variance of allele frequencies within and between populations by subtracting the heterozygosity in a subpopulation from the heterozygosity

in the total population and standardizing by the total heterozygosity (Lewontin & Krakauer 1973). In a population without substructure, where allele frequencies are evenly spread between individuals, F_{ST} equals zero. Conversely, when two populations are maximally subdivided and different alleles are fixed in each, the F_{ST} value of the metapopulation will be 1 (Holsinger & Weir 2009). A local directional selective pressure can increase the frequency of a variant in one population but not another, shifting the F_{ST} value at the selected site closer to 1 (Holsinger & Weir 2009).

1.4.1.1. F_{ST}: Time depth of selection

F_{ST} scans detect selection events 50-75kya, or 2-3 thousand generations ago (Sabeti et al. 2006). The method has power to detect selection during the final stages of a sweep, when alleles approach fixation and become highly diverged (Vatsiou et al. 2016b). The method has been found to have over 98% power to identify selective sweeps under multiple simulated conditions (Colonna et al. 2014).

1.4.1.2. F_{ST}: Considerations & confounders

Results of F_{ST} scans are affected by the choice of comparative populations, which is often subjective (Duforet-Frobourg et al. 2015). Comparisons of more closely related populations will identify recent selection in a smaller geographic range (Pickrell et al. 2009).

There is high variability in individual marker scores, even for SNPs which are closely located. This is because individual SNPs have distinct genealogies, and can be highly diverged due to neutral processes. Single-locus estimates are much noisier than the average over a window, so a signature of divergent selection is instead identified as a stretch of adjacent SNPs with a high mean F_{ST} value (Weir et al. 2005; Manel et al. 2016).

This method has reduced power in the presence of hierarchical population structure, recent admixture and selection which acts similarly in both populations compared (Duforet-Frobourg et al. 2015). Background selection may increase differentiation at a locus by removing variation in one population but not in another, and resemble positive

selection (Fagny et al. 2014). F_{ST} tests for selection are more vulnerable to false negatives than false positives, which is preferable (Manel et al. 2016).

F_{ST} doesn't indicate which of the two populations compared were affected by positive selection. This could have been achieved with a three-way F_{ST} method such as Locus-specific branch lengths LSBL (Shriver et al. 2004), which relies on a third population as an outgroup to determine the proportion of differentiation attributed to each population.

1.4.2. EHH

Under neutrality, an allele takes approximately 1 million years to reach a high frequency, and during this time the surrounding haplotype will usually be fragmented by recombination (Sabeti et al. 2006). Within 30,000 years, recombination will occur at least once in a 100 kilobase (kb) region (Sabeti et al. 2006). Therefore, common alleles typically arose long ago and occur on short haplotype blocks. However, a selected allele rapidly increases in frequency and occurs on a long, homozygous haplotype which is inconsistent with neutral expectations (Sabeti et al. 2002; Chen et al. 2010). Alternative alleles at the locus occur on a haplotype which resembles genome-wide patterns and serve as controls for the local recombination rate (Liu et al 2013). While LD under neutral conditions spans less than 0.02 cM, a selected haplotype ranges over 0.25 cM, and is detectable against background LD (Sabeti et al. 2002). A selective advantage of 1% typically produces a stretch of EHH which spans ~600kb (Sabeti et al. 2006). EHH is the probability that two randomly sampled chromosomes are homozygous along the entire length between a core allele and a certain locus. In other words, it measures the extent of recombination as a function of distance from a focal allele, compared to other alleles at the core locus (Ferrer-Admetla 2014). EHH values range from 0, indicating that all regions carrying the core allele differ, to 1 when the region is homozygous in all chromosomes carrying the core allele (Sabeti et al. 2002). The integrated haplotype score (iHS) and cross-population extended haplotype homozygosity (XP-EHH) statistics, examined in this study, are derived from EHH.

1.4.2.1. EHH: Time depth of selection

LD begins to break down once the selected allele reaches an intermediate frequency, and is negligible by the time it reaches fixation, so EHH-based methods best detect the late stage of a sweep (Sabeti et al. 2002; Voight et al. 2006). These methods are underpowered to detect recently begun sweeps, or the slow-acting effect of a small selection coefficient (Cadzow et al. 2014). The signature is only detectable transiently, for about 10,000 years (400 generations) in humans, and only while selection is ongoing. Given the selection coefficients typical for humans, most advantageous variants that arose since the Bantu dispersals haven't reached fixation. These recent variants are likely to remain polymorphic under the influence of changing environmental pressures, so EHH is a suitable approach to detecting selection in the human genome (Sabeti et al 2002).

1.4.2.2. EHH: Considerations & confounders

Recombination rates vary throughout the genome, and recombination cold spots may mimic selection signals, while recombination hotspots can break up a long range haplotype and mask selection (Liu et al. 2013; Macleod et al. 2014). EHH-based methods rely on combined linkage-physical maps (Matise et al. 2007), which contain information on genetic distances, to prevent fine-scale LD patterns from disproportionately affecting estimates at different sites (Sabeti et al. 2006). Such maps are created from the data of multiple populations to control for population-specific demographic effects and local selection (Sabeti et al. 2007). EHH-based methods will have almost no power to detect sweeps which occurred in all populations from which the genetic map was constructed (Liu et al. 2013). Since haplotype blocks are shorter in Africa, the genetic distance between two SNPs in African genomes is biased towards over-estimation, while genetic distance is likely to be under-estimated in European populations (Liu et al. 2013). Despite this, EHH-based statistics have highest power in African populations. For example, iHS has 78.9% power to detect sweeps in African populations, and only 40.98% in Eurasian populations (Pickrell et al. 2009; Fagny et al. 2014). The statistic is reported to out-perform SFS-based measures (Sabeti et al. 2006) while also concurring with SFS detected signals (Voight et al 2006).

The effect of LD isn't masked by purifying selection, but non-equilibrium demographies can cause over-estimation of haplotype-based statistics (Pickrell et al. 2009; Fagny et al. 2014). The method has less power for soft sweeps, but maintains reasonable power if the selection coefficient is high (Fagny et al. 2014; Vatsiou et al. 2016b). Although SFS-based methods are more strongly confounded by demography, Ferrer-Admetlla et al. (2014) showed that population size changes also affect haplotype-based methods, in both European and African populations. A severe bottleneck can reduce rare variants and contribute to false positives in EHH-based selection scans which detect an increase in homozygosity (Pavlidis et al. 2012). Admixture may also create a false positive signal if an extended haplotype is introduced from another population. Alternatively, migrations can introduce variation which dilutes the pattern of homozygosity produced by selection (Vatsiou et al. 2016b). Genotype call errors can reduce the sensitivity of EHH-based statistics, as the miscalled variants can cause extended haplotypes to erode (Fagny et al. 2014). The iHS and XP-EHH methods introduced below are based on the EHH statistic, so the properties discussed above also apply to sections 1.4.2.3 and 1.4.2.4.

1.4.2.3. iHS

The iHS statistic builds on EHH by comparing the EHH of the ancestral and derived alleles in a single population. The alternative allele is assumed to be unaffected by selection, so serves as a control for background LD structure (Cadzow et al. 2014). iHS is computed by finding the integral of EHH along the chromosome, in both directions away from the core SNP, until EHH reaches a value of 0.05, or a distance of 2.5 mega base pairs (Mb) from the core SNP. The use of integrals accounts for both high EHH over a short range and moderate EHH over an extensive distance. iHS is calculated for both the ancestral allele (iHS_A) and the derived allele (iHS_D) and the unstandardized score is the natural logarithm of the ratio of these scores: $iHS=\ell n(iHS_A/ iHS_D)$ (Liu et al. 2013).

Selection most often acts on the derived allele, producing a large negative iHS score, but selection on an ancestral allele is also possible, producing a large positive score (Sabeti et al. 2006). In order for iHS values to be comparable for SNPs of different allele frequencies, the score is normalized with respect to the mean and standard deviation of SNPs with a similar allele frequency to the core SNP (Sabeti et al. 2006).

An iHS value is calculated for each SNP in a dataset with a minor allele frequency (MAF) >5%, considering it the core SNP. Since a selective sweep affects the alleles around the target of selection, a collection of closely located SNPs with extreme iHS scores will produce a less variable signal than individual SNP scores (Manel et al. 2016). Although each class of methods is susceptible to heterogeneity in recombination rates, EHH-based methods using a ratio of alternative alleles are most robust, especially under the effect of population structure (Sabeti et al. 2006). The use of ratios can also control for mutation rate variation and model misspecifications (Ferrer-Admetla et al. 2014). iHS is powered to detect sweeps in samples of at least 40 chromosomes (Pickrell et al. 2009).

iHS has almost no power to detect a selected allele at low frequencies, but power reaches 80-100% at a frequency of 40% (Fagny et al. 2014). The statistic reaches maximum power for allele frequencies of 50-80%. iHS has little power to detect sweeps which are approaching or have reached fixation because there are then few alternative alleles in the population with which to compare the selected allele (Sabeti et al. 2007; Vatsiou et al. 2016a).

1.4.2.4. XP-EHH

XP-EHH is an extension of iHS which takes into account population differentiation. This method detects sweeps which have approached or reached fixation in one population, while the core SNP remains polymorphic in the comparative population. (Sabeti et al. 2007; Vatsiou et al. 2016a). XP-EHH contrasts iHS in two populations by calculating the score separately for each, and then finding the ratio of scores.

EHH is calculated from a core SNP to a given distance away for all chromosomes in population A, before the result is integrated and called I_A. I_B is calculated in the same way for population B. XP-EHH is the log of the ratio of the integrated EHH in the two populations: XP-EHH= $\ell n(I_A/I_B)$. A region with a high density of extreme values constitutes evidence of selection, with positive values of the statistic indicating selection in population A, and a negative values pointing to selection in population B. As with iHS, the genome-wide score distribution is normalized to have a mean of zero and a unit

variance (Sabeti et al. 2007). XP-EHH maintains power with sample sizes of at least 20 chromosomes (Pickrell et al. 2009).

XP-EHH is a complementary approach to iHS, as inter-population comparisons may reveal signals too weak to detect within a population. While iHS detects sweeps at intermediate frequencies and loses power as the sweep approaches fixation, XP-EHH reaches its maximum power for allele frequencies of 80-100% (Voight et al. 2006; Vatsiou et al. 2016a). Neither method has power to detect sweeps at low frequencies (<30%).

1.4.3. Tajima's D

The site frequency spectrum quantifies the derived variants in a genomic region and represents the proportion of alleles at various frequencies. As an allele sweeps to fixation, the local depletion of variation causes a population-wide shift in the SFS away from intermediate-frequency alleles (Lewontin & Krauker 1973; Cadzow et al. 2014). The Tajima's D statistic measures this shift by comparing the number of pairwise differences between individuals to the number of segregating sites, which is equivalent to comparing the amount of diversity observed to the diversity expected in the absence of selection (Lewontin & Krauker 1973). It is calculated by subtracting Watterson's estimator from Tajima's estimator and dividing the result by the variance between the two estimators (Tajima 1989). A selective sweep initially causes a reduction in variation around the selected allele, as a single haplotype begins to sweep to fixation. This results in a local negative value of D. As the allele reaches fixation, new mutations appear in the homogenous background, creating an excess of rare variants. This inflates the number of segregating sites, in comparison to the number of pairwise differences and contributes to the negative value of D (Braverman et al. 1995).

1.4.3.1. Tajima's D: Time depth of selection

Tajima's D statistic has power to identify selection only during the later stages of a sweep and briefly after fixation (Zhai et al. 2009). SFS-based methods have greater power to detect older selective events, as they make use of the new mutations which accrue over time on the haplotype containing the selected allele (Cadzow et al. 2014).

SFS-based methods under-perform compared to EHH-based methods when the frequency of the selected allele is below 90%. Above this frequency, SFS-based methods have comparable power to EHH-based methods, and reach their maximum power when the allele has recently reached fixation (Ferrer-Admetlla et al. 2014). This corresponds to sweeps within the last 250,000 years, or 10,000 generations (Sabeti et al., 2006).

1.4.3.2. Tajima's D: Considerations & confounders

Selection tests based on allele frequency distributions are most susceptible to population size changes, which alter the SFS (Tajima 1989; Cadzow et al 2014; Bank et al. 2014). Population expansions increase the proportion of rare variants, mimicking a selective sweep, while bottlenecks deplete genetic diversity and may mask a signature of selection (Nielsen 2005). Known population expansions such as the Bantu expansion and population growth of many African populations spurred by agriculturalism (Tishkoff et al. 2009; Campbell et al. 2014) might thus have created genomic patterns resembling signatures of selection.

1.5. Population-specificity of selection

Patterns of variation are known to differ between African and non-African genomes, and these differences can impact selection studies. A number of studies have found that selection is less prevalent in African than non-African genomes (Voight et al. 2006; Coop et al. 2009; Granka et al. 2012; Liu et al. 2013; Colonna et al. 2014). The following features specific to African genomes may contribute to distorting genomic signatures of selection when selection truly takes place. A higher proportion of soft sweeps has been observed in African populations (Schrider & Kern 2017), and this form of selection may be missed by methods which primarily detect hard sweeps. Additionally, regions of EHH are on average shorter in African populations, and less likely to be detected (Voight et al. 2006).

A number of studies have shown that most signals in Africa are unique to a single population, but significant sharing between populations also occurs (Liu et al. 2013; Schrider & Kern 2016). Corresponding to this finding, regions with an extreme iHS score

in one population, but not in a second generally have a high F_{ST} value (Sabeti et al. 2006). Liu et al. (2013) found that signals tend to be common to populations with shared ancestry, and that this occurs due to inheritance of the haplotypes rather than convergent evolution.

1.6. Concordance between selection study results

There is little agreement between lists of selection candidate genes produced by different studies of the same populations, and only a small proportion of selection candidates have been replicated (Voight et al. 2006; Hernandez et al. 2007; Oleksyk et al. 2010; Fagny et al 2014). For example, Lachance & Tishkoff (2013) noted that there is little overlap between the results of two genome-wide selection studies on the Pygmy, Hadza and Sandawe populations (Granka et al. 2012; Jarvis et al. 2012). Another meta-study by Haasl et al. (2016) examined various studies on genetic adaptation to high altitude and observed that Bigham et al. (2010) identified the *EGLN1* locus as a putative target of selection in the Tibetan and Andean populations, while Eichstadt et al. (2014) did not. Both reviews, together with others (Kelley et al. 2006; Manel et al. 2016), have drawn attention to the population specificity of signatures of selection, as well as discordant results from independent studies on a single population.

This low concordance could be attributed to the use of different methods, which have dissimilar statistical properties and are purposed to detect selection at distinct time depths or stages of a sweep. Studies only report their most significant results, and differences in the empirical outlier thresholds used may strongly influence results (Pavlidis et al. 2012; Fagny et al 2014). Ascertainment bias is another potential source of discrepancy in selection studies.

1.7. Ascertainment bias introduced by SNP genotyping arrays

SNP density and the distribution of markers across the genome can affect the accuracy of selection scans. Ascertainment bias can occur when the SNP markers constituting a genotyping array are not representative of the full complement of genetic variation

across the genome, resulting in parameters deviating from expected values (Nielsen et al 2005). This bias is introduced in the two-phase sampling process whereby SNPs discovered by resequencing a typically small number of individuals are used to genotype a larger sample of individuals (Weir et al. 2005; Albrechtsen et al. 2010). The probability of discovering a SNP is directly related to its frequency, so rare alleles are likely to remain undiscovered, while SNPs at intermediate frequency will be over-represented (Sethupathy & Hannenhalli 2008). Most SNP arrays only represent polymorphisms with MAF >5%, and have higher average derived allele frequencies than WGS data (Nielsen et al. 2011; Macleod et al. 2014). As a result, the frequency distribution is distorted and statistics relying on it are likely to be inaccurate (Granka et al. 2012; Clark et al. 2005). The marker SNP density of arrays is uneven across the genome, with various arrays either over- or under-representing different regions.

1.7.1. The effect of ascertainment bias on selection summary statistics

Statistics which detect the distortion in the SFS caused by selection are expected to be most strongly affected by ascertainment bias (Chen et al. 2010; Ferrer-Admetlla 2014). The increased proportion of common variants in arrays is likely to inflate Tajima's D measures and mask signatures of selection (Clark et al. 2005; Voight et al 2006). SFS-based statistics were designed for full-sequence data, but have been used on ascertained data with the hope that genomic regions in which the SFS deviates the most will be identified nonetheless (Voight et al 2006). The bias towards high-frequency alleles was a major concern to studies of the HapMap data, and Sabeti et al. (2006) avoided test statistics relying on derived allele frequencies because of this.

F_{ST} is likely to be affected by ascertainment bias since it is a function of the frequency spectrum. Population-specific variation generally occurs at lower frequency, so ascertained markers will be biased towards low F_{ST} values for two populations which are both distantly related to the ascertainment populations (Clark et al. 2005; Weir et al. 2005; Chen et al. 2010). However, markers in LD with the selected variant should be highly differentiated, allowing the signal to be detected (Pickrell et al. 2009). Conversely, if a distantly related population is compared to the populations in the ascertainment

panel, F_{ST} is expected to increase (Albrechtsen et al. 2010). The extent to which F_{ST} measures are affected is determined by the differentiation between the genotyped and the ascertainment populations, so F_{ST} will be biased especially when the SNP discovery sample isn't ethnically diverse.

EHH-based methods are expected to be affected minimally by ascertainment, since markers occur on the same long haplotype as the selected variant, and should have risen to high frequency together with it (Sabeti et al. 2006). However, LD estimates might be either increased (Macleod et al. 2014) or decreased by low SNP density (Clark et al. 2005; Macleod et al. 2014; Goodwin et al. 2016). Since tag SNPs are chosen across the genome to represent haplotype blocks, areas with high LD are likely to be more sparsely sampled. Pickrell et al. (2009) analyzed data produced by the Human Genome Diversity Panel containing 650,000 common SNPs and found that the array has fewer markers in selected regions (as detected in HapMap data) than the genome-wide average. The effect of ascertainment bias is exacerbated when SNPs are ascertained from a population distantly related to the genotyped population, and coalescent simulations have shown that lower LD estimates occur as a result (Chimusa et al. 2015). Additionally, the iHS signal-to-noise ratio is lower in WGS data sets than in genotyping data sets because extended haplotypes are more rapidly broken in the presence of low-frequency variants (Grossman et al. 2013).

1.7.2. Correction for ascertainment bias in selection stiudies

Bias could be diminished by modelling the ascertainment process, but this requires knowledge of the SNP discovery protocol used to create the genotyping platform, and this information is often unavailable (Ramírez-Soriano & Nielsen 2009; Chen et al. 2010). The ascertainment process is complex and perhaps impossible to replicate (Nielsen et al. 2005; Lachance & Tishkoff 2013; Macleod et al. 2014). For example, the Perlegen study removed population identifiers, while the HapMap (International HapMap Consortium 2003) ascertainment criteria changed during the study. Clark et al. (2005) compared the heterozygosity of HapMap and Perlegen array data and found that the metrics differed significantly for the two datasets, even after ascertainment correction. They found that the HapMap data had a higher heterozygosity (a higher percentage of

common SNPs), and reasoned that the larger size of the Perlegen discovery sample would have facilitated identification of rarer alleles. This is expected, since the HapMap array was designed to capture common variation, but demonstrates the effect of the ascertainment process to skew genome-wide summary statistics. Correction for ascertainment relies on accounting for the ancestral constitution of the ascertainment sample but most SNP selection procedures rely on data from dbSNP (Sherry et al. 2001), which was compiled from studies using different sample sizes and unclear ethnic distributions (Albrechtsen et al. 2010).

1.7.3. Ascertainment bias in Africa

More prevalent alleles are likely to be older and present in multiple populations, so a shift in the SFS caused by ascertainment bias may under-represent population differentiation. Since there is high divergence among some African populations, different groups might poorly represent the diversity of others (Gurdasani et al. 2015). Very rare variants may constitute a large proportion of African genomes, underlying the 'missing variability' that has been observed by numerous studies (Macleod et al. 2014). Since diversity varies both within and among populations, bias can be introduced when an array is used to genotype individuals from a population other than the one in which the markers were ascertained. This is especially problematic in African studies, as most arrays are Eurocentric and underrepresent African diversity (Colonna et al. 2014). The SFS differs in populations with African ancestry, and a greater proportion of selection candidates occur at lower frequencies of 30% or less. The abundance of rarer alleles corresponds to a genome-wide reduction in F_{ST}, indicating increased genomic diversity (Liu et al. 2013). Additionally, due to shorter haplotypes in Africa, a higher density of tag SNPs is required to represent African genomes (Chimusa et al. 2015). Some (but not all) studies have indicated that selection is less prevalent in African populations (Colonna et al. 2014; Granka et al. 2012). This observation may have been confounded by the use of Euro-centric genotyping arrays, since ascertainment has been less thorough in African populations (Colonna et al. 2014).

For example, Granka et al. (2012) studied selection in Africa and found that patterns of SFS differentiation and haplotype sharing between populations were consistent with

neutral expectations, implying that many outliers were false positives. They suggested that population history is more causative of allele frequency differentiation than selection, since the extent of candidate overlap between populations mirrored genome-wide patterns produced by population histories. They attributed this partially to ascertainment of SNPs from European populations, reasoning that the markers are unlikely to be in LD with SNPs which are selected in African populations.

1.7.4. Comparing identified selection signals between genotyping array and WGS data

Until recently, selection studies have primarily been performed on array data, and investigators have speculated that the density of ascertained SNPs may affect statistics, even when ascertained from a closely related population (Sabeti et al. 2006; Granka et al. 2012). WGS data has allowed this unbiased characterization of the spectrum of allele frequencies in the genome, and has revealed much previously unknown diversity. For example, Phase 1 of the 1000 Genomes Project (1000 Genomes Project Consortium 2010) provided ten times more information than the HapMap Project (International HapMap Consortium 2003; Fagny et al. 2014) and this increase in SNP density is expected to improve the power to detect signatures of selection (Fagny et al. 2014). This warrants comparative investigation of the results of WGS and genotyping array data. Although the availability of WGS data is increasing, the use of array data has continued in recent studies, so such a comparison is relevant both in interpreting results of previous studies and in the choice of data to use in future studies.

1.8. Study rationale

Due to the scarcity of WGS data, genotyping array data has been analyzed by many selection studies. Arrays have typically been designed to represent more common alleles in association studies, and have been shown to skew population genetic summary statistics (Nielsen et al 2005; Albrechtsen et al. 2010). Additionally, arrays are known to represent African genomes less effectively than the European populations (Colonna et al. 2014; Chimusa et al. 2015), which are over-represented in ascertainment samples. A direct comparison of the results of selection scans from different genotyping

arrays to results of WGS data has not yet been conducted. Low levels of overlap have been reported for different selection studies (Voight et al. 2006; Hernandez et al. 2007; Oleksyk et al. 2010; Fagny et al 2014) and multiple variables including ascertainment bias may have contributed to differences in results. For example, various selection statistics may be differentially affected by reduced SNP density. Additionally, if population structure is present, ascertainment bias may be introduced on the level of the individuals sampled to represent a population.

1.8.1. Aim

To determine the effect of ascertainment bias and other potential influences on the results of selection studies.

1.8.2. Objectives

1. Assess the accuracy with which various genotyping arrays represent whole genome sequence data in selection studies and identify the arrays best suited for African populations
2. Examine the effect of SNP density per window on the detection of selection signals
3. Assess the concordance between selection signal results from two independent samples of the SEB group
4. Determine the concordance between the selection signals detected by different methods
5. Quantify the amount of selection signal sharing between populations
6. Annotate selection candidates with gene names and functional descriptions, and highlight novel signals

2. Methods

2.1. Genetic datasets

This study analyzed publicly accessible WGS and genotyping array data from two projects, for which individual identifiers have been anonymized. Firstly, the AGVP (Gurdasani et al. 2015) sequenced individuals across sub-Saharan Africa at 3x coverage. The sampled populations include the Zulu population from South Africa, the Baganda population of Uganda, and four Ethiopian ethnic groups: the Wolayta, Oromo, Somali and Amhara. The Ethiopian sub-populations were pooled to increase sample size.

The geographic locations of these populations are shown in Figure 2.1.

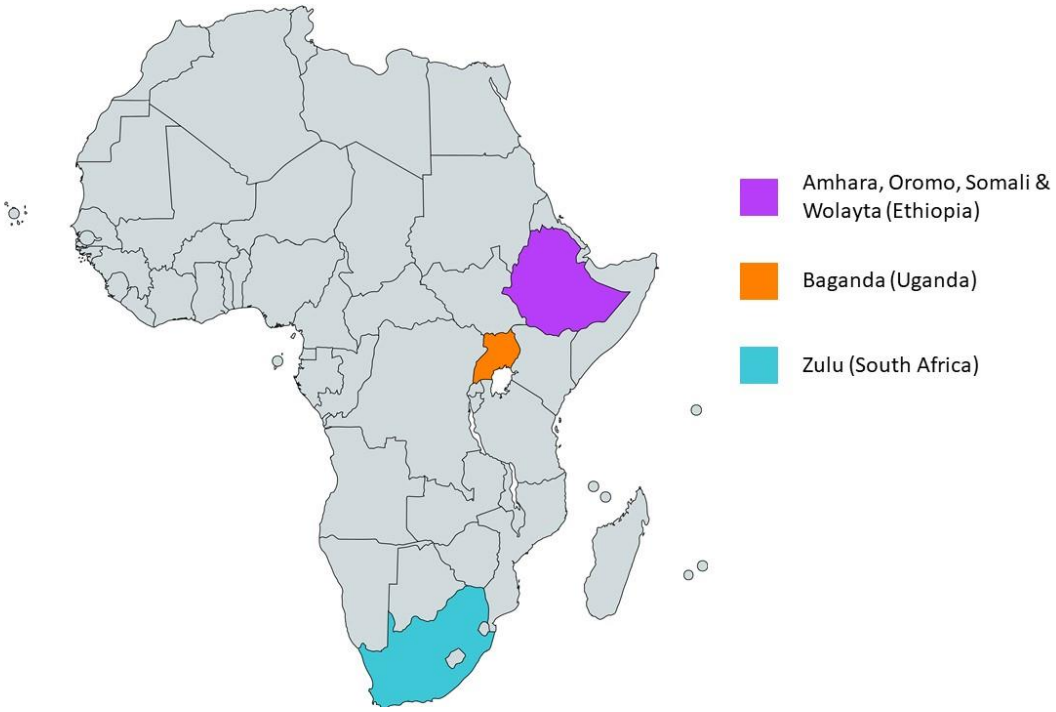

Figure 2.1: Map of Africa highlighting the locations of the populations sampled by the AGVP.

Secondly, the Birth to Twenty (Bt20) is a multidisciplinary, longitudinal study which has included genotyping a subset of the Sowetan cohort with the high-density Illumina Human Omni 5 genotyping array (May et al. 2013). The population of Soweto, South

Africa, constitutes ethnolinguistic groups belonging to the broader group of SEB group. The sample subsets are summarized in Table 2.1.

Table 2.1: Samples included in this study

Population	Country	Data type	Study	Sample size
South African Black (Zulu-speaking)	South Africa	WGS	AGVP	100
Baganda	Uganda	WGS	AGVP	100
Amhara, Oromo, Somali & Wolayta	Ethiopia	WGS	AGVP	72
South African Black	South Africa	Illumina HumanOmni 5M	Bt20	94

2.2. Raw data quality control

Data was received in the form of Variant Call Format (VCF) files, which had previously been filtered to retain only high quality variant calls. Quality control was performed for autosomal data with PLINK version 1.9 (Purcell et al. 2007) and VCFtools version 0.1.16 (Danecek et al. 2011) software. This involved removing indels, as well as multi-allelic and duplicate SNPs. Individuals with missingness >2% and SNPs with missingness >1% were removed. No MAF cutoff was applied to the input data for F_{ST} and Tajima's D. Although it is advised to remove low-frequency alleles in low coverage WGS data, these selection statistics rely on rare alleles to identify signatures of selection. Conversely, EHH-based scans are concerned with alleles which have risen to appreciable frequency, so for iHS and XP-EHH, alleles with a frequency <5% were automatically excluded by Selscan (Szpiech et al. 2014). As EHH-based scans use genetic distance information to control for genomic recombination rate variation, genetic map coordinates were added from the Rutgers Combined Linkage-Physical Map (Matise et al. 2007). Phasing was performed for the Bt20 data with the 1000 Genomes Phase 3 data (1000 Genomes Project Consortium 2012) as a reference, using SHAPEIT2 (Delaneau et al., 2013). The AGVP data was previously phased using SHAPEIT2 and standard parameters (Gurdasani et al. 2015 supplemental information) with the 1000 Genomes panel (1000 Genomes Project Consortium 2010) as a reference.

2.3. Generation of *in silico* genotyping data

The effect of ascertainment bias in selection studies was assessed by comparing the results of selection scans from WGS data to those of array data. The WGS data was downsized based on the SNP contents of different genotyping arrays, to generate *in silico* genotype data, using VCFtools. Data produced by real genotyping arrays has lower genotyping error rates than low coverage WGS data, so the process followed here might not perfectly represent real array data (Nielsen et al. 2011). This potential discrepancy would be expected to most strongly impact Tajima's D, which relies on rare alleles to detect selection. The genotyping arrays examined in this study include some of the most widely used high-density genome-wide arrays, as well as two which aim to represent African diversity. The arrays are summarized in Table 2.2. The Axiom® Genome-Wide PanAFR Array Set (PanAFR) was designed to capture the genetic variation of the Yoruba population, while the H3Africa Array (H3A) was recently designed to capture diversity in numerous populations of sub-Saharan Africa. Many marker SNPs of the Illumina Omni 1 and Omni 2.5 arrays are contained in Omni 5, and all three arrays were assessed to investigate whether some signals of selection are missed when using a less dense panel. The list of SNPs included in these arrays was provided by thhe H3Africa array design team.

Table 2.2: SNP markers (millions) per array, before and after quality control (QC)

Genotyping Array	Markers pre-QC	Markers post-QC
Affymetrix 6.0 (Affy 6)	0.90	0.88
Illumina Omni 1.0 (Omni 1)	1.10	0.96
Mega 2.5	1.40	1.35
Illumina H3A (H3A)	2.30	2.07
Affymetrix PanAFR	2.20	2.15
Illumina Omni 2.5 (Omni	2.30	2.18
Illumina Omni 5 (Omni 5)	4.30	3.67

2.4. Methods used to detect selection signals

Four commonly used scans for signatures of selection were performed on both the WGS and simulated genotyping array data.

Weir and Cockerham's F_{ST} statistic (Weir & Cockerham 1984) was computed for each pair of populations with VCFtools. The input consisted of a VCF file containing both populations compared, together with a text file containing individual identifiers per population. Tajima's D statistic was also obtained using VCFtools. The program requires a separate input file for each population, containing standard genotype data in VCF format. iHS and XP-EHH scans were performed with Selscan version 1.2.0 (Szpiech 2014). The program requires phased, biallelic genotype data in VCF format, as well as a PLINK formatted map file composed of physical and genetic map information. XP-EHH requires a separate input VCF file per population, and the two files must contain identical loci. The default Selscan parameters were used, in accordance with Voight et al. (2006). Scores were standardized by allele frequencies, using the program norm version 1.2.1 provided by GNU GSL.

A signature of selection is formed by a region of SNPs with extreme values for a selection statistic, and single SNP scores are too variable to be considered as signals. Thus, scores of all SNPs within a 10kb window were summarized (Manel et al. 2016).

For F_{ST} scans, the mean weighted F_{ST} for the region was estimated by VCFtools, whereas for EHH-based scans the average of the highest 5 scores in a window was considered. Non-overlapping windows were used to facilitate comparison between methods, populations and arrays. Windows in the iHS and XP-EHH results were retained only if they contained at least 20 SNPs with selections statistic scores for WGS data and 10 SNPs for array data. The cut-off was lowered for array data to compensate for reduced SNP density. From the results of F_{ST} and Tajima's D scans, windows were discarded if they contained fewer than 10 SNPs per window.

Selection signals were identified as outliers: windows with the most extreme scores for the selection summary statistics. Windows were sorted according to their value for each statistic and regions beyond a certain percentile of the score distribution were then

isolated.This threshold can be considered the minimum empirical P-value for the windows in the score distribution (Teshima et al. 2006).

A signature of selection usually spans a large region, so adjacent outlying windows are likely to be part of the same signal. Therefore, outlying windows separated by a distance of 10kb or less were merged using BEDtools version 2.14.3 (Quinlan & Hall 2010) before further comparison.

2.5. The accuracy with which various genotyping arrays represent whole genome sequence data in selection studies

The concordance between array and WGS results was assessed by assigning accuracy measure classifications to each genomic window. The WGS outliers were considered the 'true' results, while the array outliers were labelled as 'positive' results. True positive (TP) results were identified as windows which were outliers in the score distributions of both array selection scan results and the WGS results. Windows were considered false positives (FP) if they were outliers in the results of an array but not the WGS score distribution. False negative (FN) windows were those which were outlying in the WGS score distribution, but not in the results of an array. Finally, true negative (TN) windows were defined as regions which fell below the outlier threshold in both the WGS and array results. These classifications are summarized in Table 2.3.

Table 2.3: Contingency table for the accuracy of selection signals identified from array data with various outlier thresholds, in comparison to WGS 1% outliers.

	WGS top 1% outliers (selected)	WGS 99th percentile (not selected)
Array top *x*% outliers (selected)	TP	FP
Array (100-*x*)th percentile (not selected)	FN	TN

Since adjacent outlying windows had been merged, windows were not equally sized and could not be compared directly between the WGS and array data by their start positions. Instead, the overlap of windows was considered when comparing datasets. BEDtools was used to find the various intersections between datasets. The true positive rate (TPR) and false positive rate (FPR) for 1% outliers were represented per array in bar graphs per population and method. The TPR was computed by dividing the number of TP windows by the sum of the TP and FN windows: TPR=TP/(TP+FN). The FPR was calculated by dividing the number of false positive windows by the sum of FP and TN windows: FPR=FP/(FP+TN). The choice of outlier threshold is somewhat arbitrary, but 1% is a commonly used cut-off.

Next, the rank of FN windows within the array score distribution was explored. It was assumed that if a WGS outlier was not within the 1% outlier threshold of an array, it could occur within a slightly wider outlier threshold. To examine the rank of these FN windows in the array score distribution, the outlier threshold for the array results were incrementally increased, and windows were classified into accuracy measure categories by comparison to the WGS 1% outliers. Various outlier thresholds were defined, ranging from 0.5% to 5%, in increments of 0.5%. These results were summarized by calculating the true positive rate (TPR) and false positive rate (FPR) for each array at each outlier threshold. The TPR and FPR over the range of outlier thresholds were visualized in a joint line graph where each array was represented as a line.

The correlation between the scores of the WGS outliers and the scores of the corresponding windows in the array results was examined. To determine the appropriate correlation coefficient, normality tests were performed on the score distributions for a few of the datasets. The Anderson-Darling test was performed for the WGS outlier score distribution, including only windows represented by the Omni 5 array. This was done for all selection summary statistics and populations at significance levels of 1, 2.5, 5, 10 and 15. The statistic scores and critical values are provided in Supplemental Table 2. All tested datasets deviated from normality, at all significance levels. This informed the choice of a non-parametric test, Kendall's Tau coefficient (also called Kendall's rank correlation coefficient). This statistics is a measure of the nonlinear relationship, or rank correlation between two distributions, on a scale of -1 to 1. For each window in the outlying 1% of the WGS score distribution, the WGS score was compared to the array

score, regardless of whether the window was an outlier in the array score distribution. A value of Kendall's tau was obtained for each data subset, defined by array, population and selection method. All the above statistical tests were implemented with Python's SciPy library version 0.19.0.

To visualize the relationship between the values of the selection test statistics of the WGS and array data, the corresponding scores of each window were plotted against each other in a scatter plot. The axes were given the same scale to enable comparison of WGS and array scores.

As a final assessment of the performance of array data in selection studies, the percentage of array outliers which were true positives was considered. The percentage of true positives (TP%) was calculated by dividing the number of true positive windows by the total number of outlying windows in the array score distribution: TP% = [TP/(TP+FP)]*100. The various arrays have different sizes, and therefore different numbers of outlying windows. The TP% can be used to compare arrays of different sizes, while comparing the absolute number of TP windows between arrays would not be informative. The TP% of the arrays were plotted in bar graphs with Python's Pandas library. Arrays were presented in ascending order of marker density.

The effect of ascertainment bias on selection studies was further explored by comparing the concordance between results of different arrays. Differences in results between arrays could be due to either unequal SNP density, or the choice of markers with which an array was created. The number of selection signals (1% outliers) shared by various combinations of arrays was found with BEDtools. To visualize the size of the intersections between the outliers of different arrays, UpSet plots were created with the R UpSetR library (Conway et al. 2017). This was done For F_{ST}, the Bt20 and Zulu samples were each paired with the AVGP for each population and selection scan method. UpSet plots display the sizes of intersections, with a group shown by dark circles below the plot, and the size of an intersection given by the height of a bar.

2.6. The effect of SNP density per window on the detection of selection signals

To assess the relationship between SNP density and accuracy measure classifications, SNP density per 10kb window in the array data was examined as a percentage of SNPs in the corresponding WGS window. Percentages, rather than absolute SNP counts, were considered to control for the variation of SNP density throughout the genome and isolate the effect of array SNP density. SNP density distributions were examined per accuracy measure category to determine how strongly these classifications are affected by SNP density. The distributions were represented with box plots, created with Python's Seaborn library version 0.9.0, and the mean, median and standard deviation were calculated for each using Python's SciPy library. To determine the significance of the difference between the distributions for each pair of accuracy measure classifications, the Mann-Whitney U statistic was calculated with SciPy.

Many windows with less than 10 or 20 scores had been removed from the results, and are not represented in these distributions. Therefore the proportion of FN windows which were removed due to insufficient SNP density was calculated.

2.7. The concordance between selection signals identified from two independent samples of the Southeastern Bantu-speaking group

To investigate the extent of overlap between signals, three selection scans were performed on two independent samples of the SEB group: the Bt20 and AGVP Zulu samples. The Bt20 sample was genotyped on the Omni 5 array, so to directly compare the AGVP Zulu sample, the SNPs corresponding to the Omni 5 array were extracted from the AGVP WGS data. F_{ST}, iHS and Tajima's D scans were performed with 10kb windows. For F_{ST}, which compares two populations, the Bt20 and AGVP Zulu samples were each paired with the AGVP Baganda and Ethiopian samples. Windows with fewer than 10 SNPs were discarded. Signatures of selection were identified as outliers in the empirical score distribution by applying three increasingly stringent outlier thresholds:

0.1%, 0.05% and 0.01%. Adjacent outlier windows were merged, before the extent of the concordance between selection signals of the Bt20 and AGVP Zulu results was assessed. Outlying windows which overlap between the two samples were identified with BEDtools version 2.27.0 and the percentage of outliers shared by both samples was calculated. The probability that two datasets of specified size share a certain number of windows was found with the hypergeometric probability formula, using the calculator on nemates.org (Lund 2005). The hypergeometric distribution was here parameterized by the total number of windows, the number of outliers and the number of outliers shared by the two datasets (Bt20 and AGVP). The representation factor was calculated by dividing the number of overlapping windows by the number expected in the intersection. A value greater than 1 indicates that the two datasets have more windows in common than expected by chance, while a value below 1 indicates less overlap than expected. A representation factor value of 1 signifies the null hypothesis and would be obtained if the number of overlapping windows matches the expectation for two independent datasets.

2.8. The agreement between the results of different methods

Signals from the four selection scans were compared for the Zulu population or Zulu and Ethiopia population pair. Single-population and two-population methods might not be directly comparable, since differences could reflect selection on the Ethiopia population. Nonetheless, the concordance was examined by finding the number of selection candidates which were identified by various combinations of methods and representing the signal sharing in an UpSet plot.

2.9. The amount of signal sharing between populations

An approach similar to Section 2.8 was employed to investigate the concordance between the signals identified in the three populations. The 1% outliers (observed for each method) shared between these populations were identified and the number of windows in each intersection of populations was represented as an UpSet plot.

2.10. Selection candidates which were not identified from any array using any selection scan method

Ascertainment bias was further explored by creating a shortlist of the strongest signals which were missed by all arrays. A stringent outlier threshold of 0.01% was applied to the results of each selection method, for both the WGS and array data. Results from the array data were compared to the WGS results and FN windows were isolated. These windows were annotated with gene names from Ensembl GRCh37 (Zerbino et al. 2017). Since signals which are identified by multiple methods are more robust, only genes which were false negatives for two selection scan methods were retained. Signals detected using methods using single-population estimates and inter-population estimates were considered separately. The resulting lists of genes consisted of the top signals which were false negatives in the results from all arrays.

2.11. Functional annotation of selection candidates

An outlier threshold of 0.05% was applied to isolate candidate regions from the WGS results. Of these outliers, only those identified by two methods per population or population pair were retained. This shortlist was annotated with gene names, gene descriptions phenotype descriptions, and gene ontology (GO) terms retrieved from the Ensembl Human Genes (GRCh37.p13) database (Ensembl Genes version 93) using the BioMart tool (Smedley et al. 2015).

The genes for which Ensembl provides both gene descriptions and phenotype descriptions were tabulated. The 50 most common GO terms across all populations and methods were identified and listed.

It was thought that this study might identify selection candidates which have not been reported by previous studies. To identify novel selection candidates, the windows in the 0.05% tail of the WGS results were compared to a modified list of previously identified selection candidates was then used for comparison (Choudhury et al. 2017 personal communications). The list was created from the dbPSHP dataset (Li et al. 2014) by removing genes from studies which listed more than 1,000 entries, since they might have used an outlier threshold which was fairly large. Genes were added from

publications post-dating 2014, which weren't included in the dbPSHP dataset. Windows of 10Kb which overlap with genes which do not appear in the modified dbPSHP database were considered as novel signals.

3. Results

3.1. The accuracy with which various genotyping arrays represent whole genome sequence variation in selection studies

The primary research question was to assess the impact of ascertainment bias of the SNP markers of commercial arrays on detecting signatures of selection. To evaluate this, the results of selection scans performed on array data were compared to the results from WGS data. This analysis was conducted for four different methods that are commonly used to assess signatures of selection to detect any possible variation in the level of accuracy between these methods. Moreover, this analysis was repeated in three different populations to verify whether the observed trends are restricted to a single population or are consistent across populations.

3.1.1. True positive rates and false positive rates of 1% outliers

Firstly, selection signals identified in array and WGS results were compared in terms of accuracy measures for each array. The 1% outliers from selection scans performed on WGS data were considered ‘true’ signals. If the same signals were also observed in the outlying 1% of the score distribution from array results, these windows were classified as ‘true positive’. Signals observed as outliers only in array results were considered ‘false positive’. The number of windows within each accuracy category was counted to calculate accuracy measures for each array, population, selection summary statistic, and outlier threshold (Appendix 2). Figure 3.1 displays the TPR and FPR for all populations.

The FPR values were much lower than the TPR values because the TN category (contributing to the denominator in the TPR equation) necessarily contained the largest proportion of windows. For all methods, TPR is correlated with the number of markers in an array. Although TPR and FPR were analyzed to partially equalize the effect of SNP density per array, it was clear that the number of markers strongly influenced the accuracy of results.

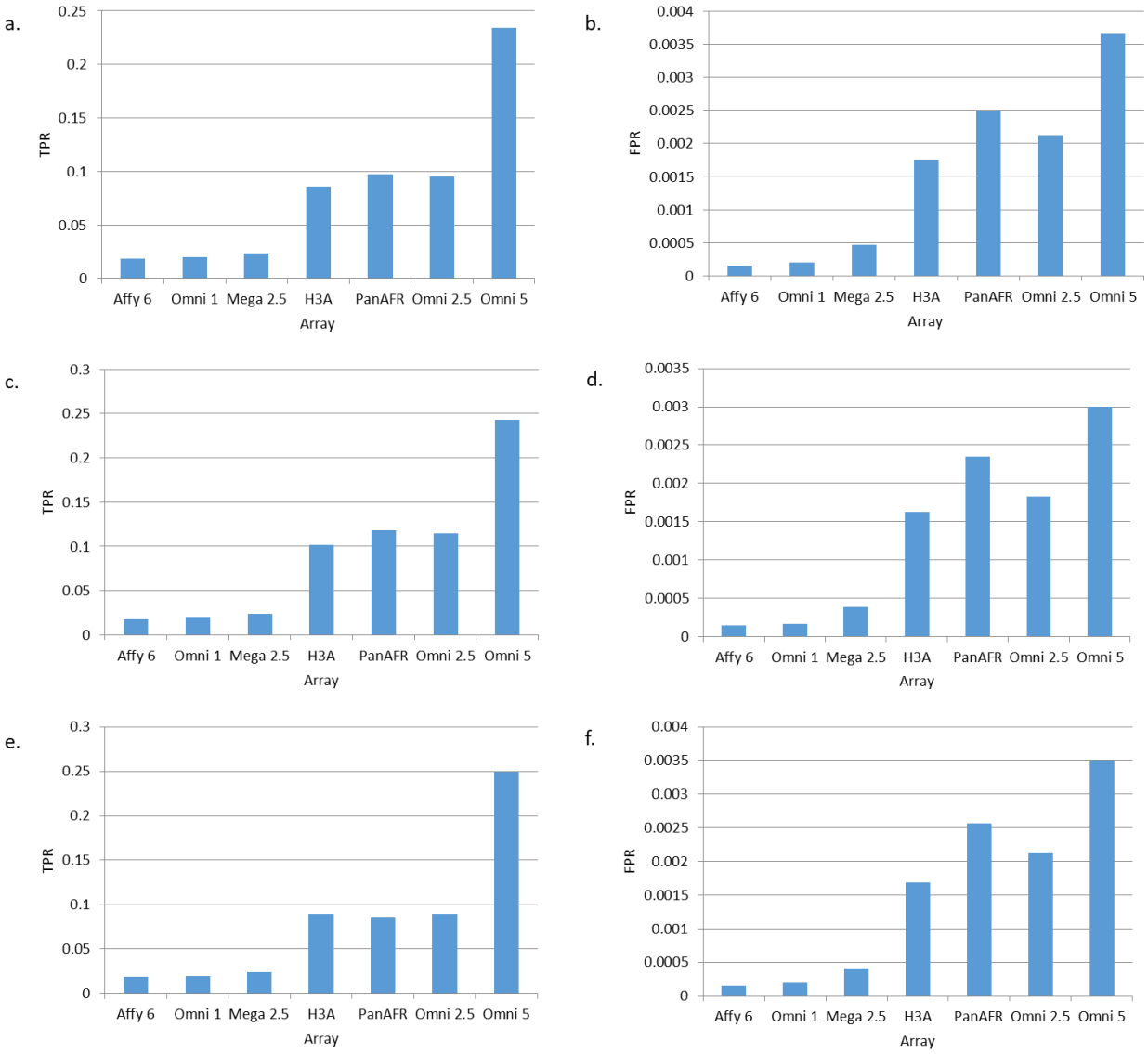

Figure 3.1.1: The true positive rate (TPR) and false positive rate (FPR) of each array at a 1% outlier threshold, for the F_{ST} method. Arrays are presented in ascending order of the number of markers they contain. a. TPR; Baganda and Ethiopia, b. FPR; Baganda and Ethiopia, c. TPR; Baganda and Zulu, d. FPR; Baganda and Zulu, e. TPR; Ethiopia and Zulu, f. FPR; Ethiopia and Zulu

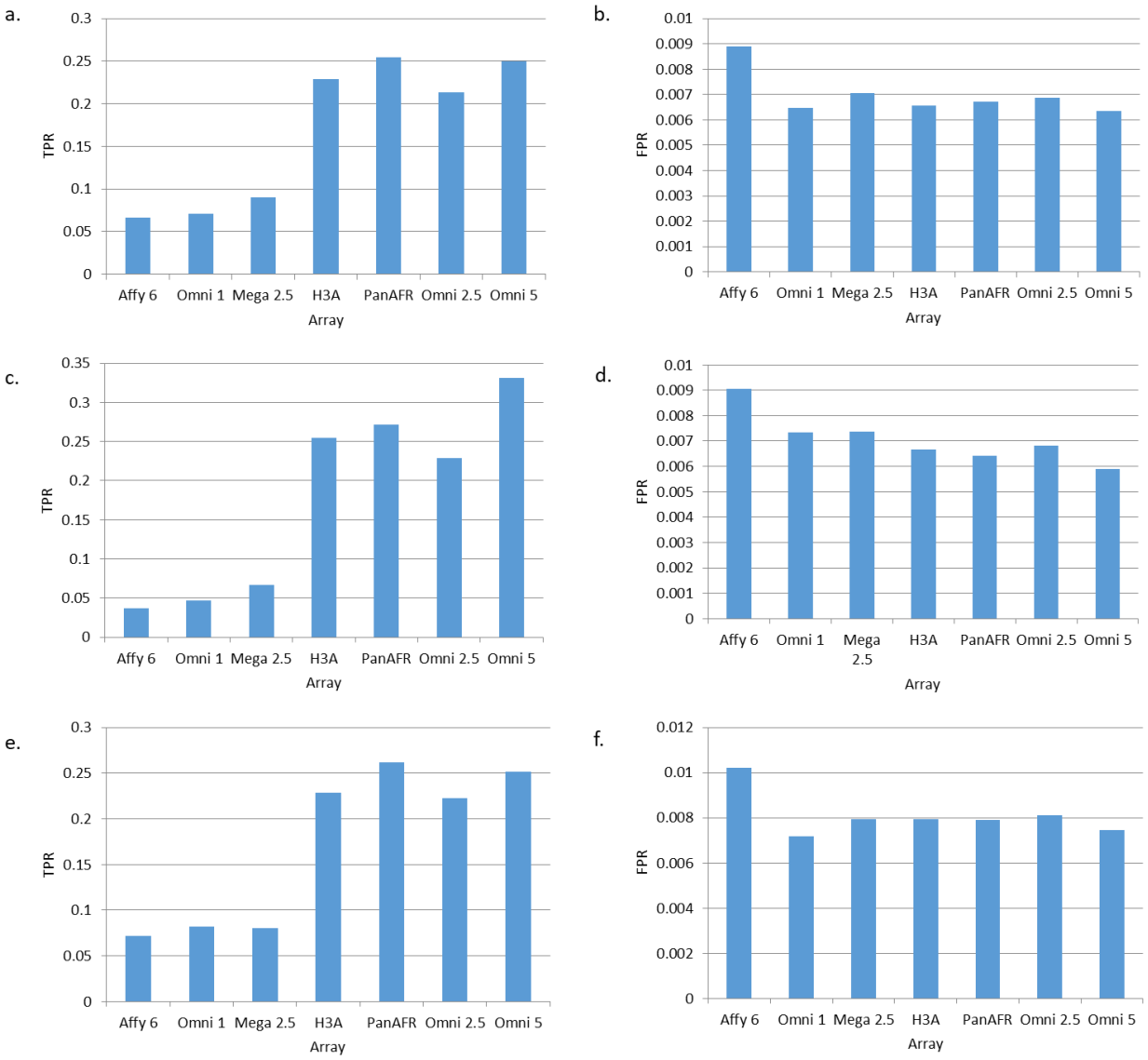

Figure 3.1.2: The true positive rate (TPR) and false positive rate (FPR) of each array at a 1% outlier threshold, for the iHS method. Arrays are presented in ascending order of the number of markers they contain. a. TPR; Baganda, b. FPR; Baganda, c. TPR; Ethiopia, d. FPR; Ethiopia, e. TPR; Zulu, f. FPR; Zulu

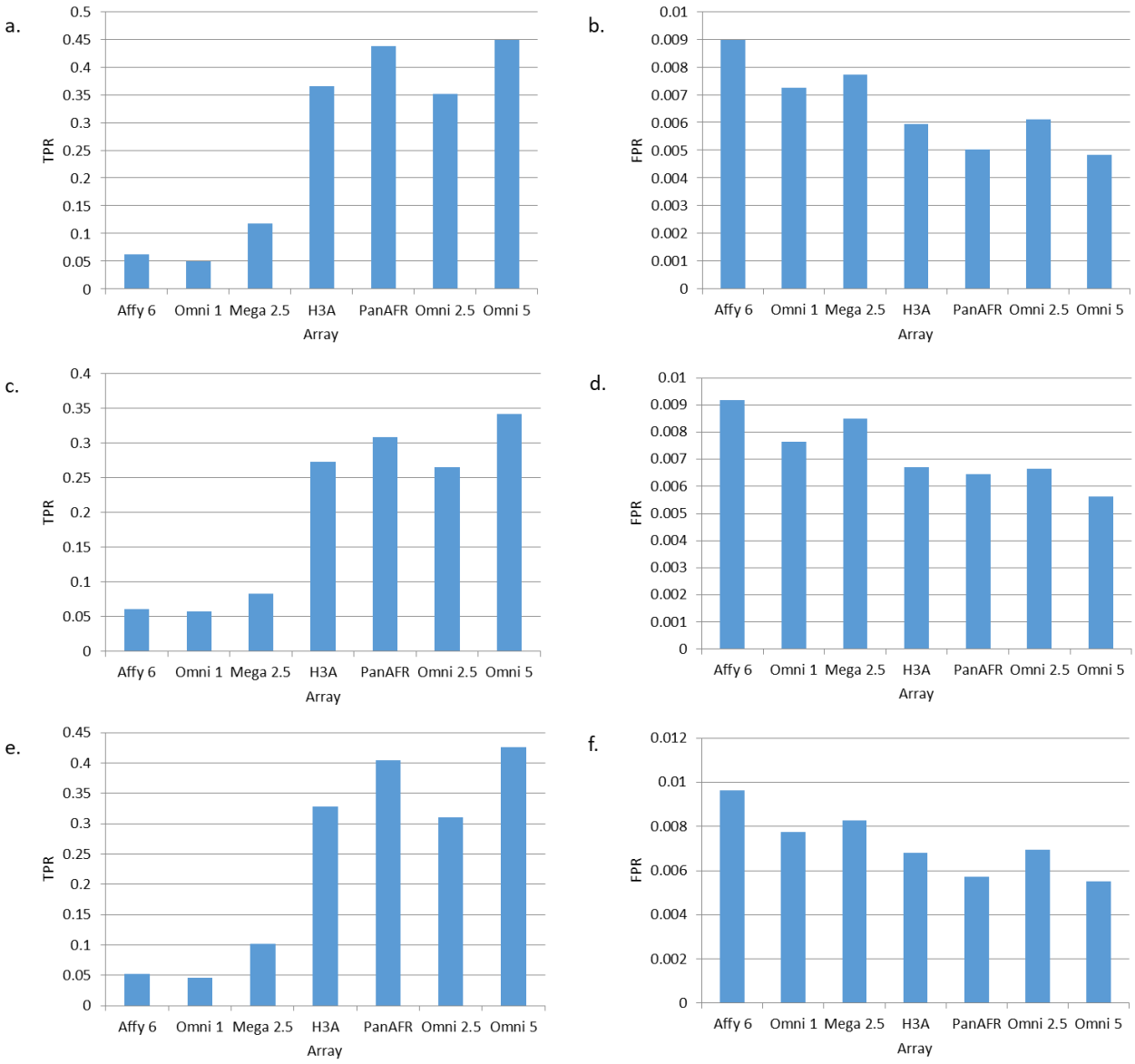

Figure 3.1.3: The true positive rate (TPR) and false positive rate (FPR) of each array at a 1% outlier threshold, for the XP-EHH method. Arrays are presented in ascending order of the number of markers they contain. a. TPR; Baganda and Ethiopia, b. FPR; Baganda and Ethiopia, c. TPR; Baganda and Zulu, d. FPR; Baganda and Zulu, e. TPR; Ethiopia and Zulu, f. FPR; Ethiopia and Zulu

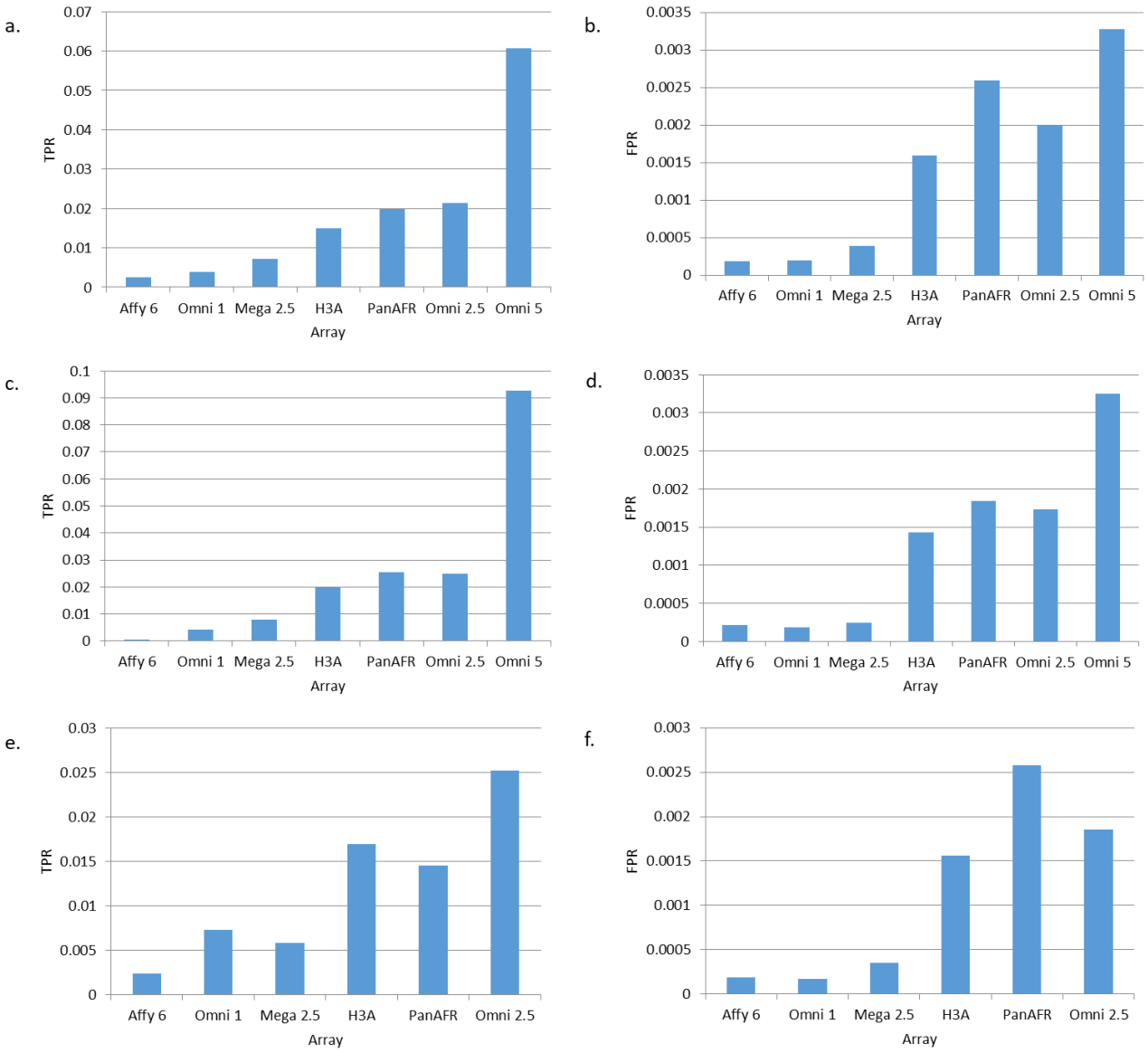

Figure 3.1.4: The true positive rate (TPR) and false positive rate (FPR) of each array at a 1% outlier threshold, for the Tajima's D method. Arrays are presented in ascending order of the number of markers they contain. a. TPR; Baganda, b. FPR; Baganda, c. TPR; Ethiopia, d. FPR; Ethiopia, e. TPR; Zulu, f. FPR; Zulu

The results of the F_{ST} method (Fig 3.1a) show that the TPR increases with array size. The arrays of approximately 1 million SNPs (M) produced TPR values of around 0.005, while the 2.5M arrays produced TPR values of around 0.08. The 5M array produced TPR values of around 0.25, similar to those of the EHH-based methods. Therefore, while the 2.5M arrays performed only 1.5 times better than the 1M array, the 5M array delivered almost 3 fold higher accuracy compared to the 2.5M array.

Results from the iHS selection scan show that higher TPR values are produced by the approximately 2.5M arrays, in comparison to the 1M arrays. However, no pronounced difference in TPR was observed for the 5M array (Omni 5), in comparison to the 2.5M arrays (Figure 3.1b). This can be explained by the array design, since the 5M array was purposed to capture variation at the level of rare variants (MAF<0.05), which were removed by the Selscan program prior to the selection scans. Conversely, the content of common SNPs (MAF>0.05) is comparable between the Omni 2.5 and Omni 5 arrays. Among the 2.5 M arrays, differences in the TPR were observed, with the PanAFR showing the highest TRP, followed by H3A and Omni 2.5, across all populations. Although differences in FPR between these arrays were less pronounced, Omni 2.5 showed the highest FPR values across almost all populations.

A similar trend was also observed for the XP-EHH estimates (Figure 3.1c). The results indicated that the two African-based arrays (PanAFR and H3A) are more accurate than Omni 2.5, a European-based array of similar size.

Tajima's D had the lowest TPR in comparison to other methods (Figure 3.1d). For example, the Affy 6 results for the Zulu data contain only 5 TP windows, leading to a low TPR of 2.42×10^{-3}. Although a trend of linear increase with size was observed, for Tajima's D, the largest array (Omni 5M) only achieved a TPR value similar to those for the smallest arrays with the EHH-based methods.

The highest TPR among these analyses was 0.45, observed for Omni 5 and the XP-EHH method for the Baganda and Ethiopian population pair (Appendix 2.4.a). Concurrently, the FPR values were highest for the smallest array and lowest for the largest array.

For all arrays and selection scan methodss, the FPR values were much lower than the TPR values. This is primarily because the TN category (contributing to the denominator in the TPR equation) necessarily contained the largest proportion of windows. For all methods, TPR was correlated with the number of markers in an array. Although TPR and FPR were analyzed to partially equalize the effect of SNP density per array, it was clear that the number of markers strongly influenced the accuracy of results.

3.1.2. The positions of WGS 1% outliers in the array score distributions

The accuracy of array results was further explored by examining the spread of WGS 1% outliers within a larger tail of the array score distributions. If the FN windows from the previous analysis had selection statistic scores which are close to the outlier threshold, this would indicate that the array delivering more accurate results than if the FN windows occurred towards the opposite tail of the distribution. The analysis was done by calculating the TPR and FPR over a range of outlier thresholds for the arrays, in comparison to the WGS 1% outliers. Figure 3.2 shows the results for the Zulu population and Ethiopia and Zulu pair. Plots for the remaining populations are available in Appendix 3.

The metrics plotted are reminiscent of a ROC curve, but these plots deviate from conventions by containing values for only the outlier thresholds of interest. As a result, the axes have different scales and the TPR and FPR ranges differ for each array. The outlier thresholds are not explicitly shown because this is a convention for ROC curves.

The results show a clear relationship between array size, ascertainment bias of the array and the algorithm used for detecting signatures of selection. No MAF filter was applied for F_{ST} and Tajima's D. As a result, the most accurate results were produced by the largest array (Omni 5), which contains the largest number of rare alleles. The difference between the 2.5M arrays and the 1M arrays was more pronounced with the F_{ST} method than with Tajima's D.

With the EHH-based methods, the larger arrays produced higher TPR values and lower FPR values. The difference between 2.5 M arrays and the 5M array was much more notable for the XP-EHH based estimates. Among the three 2.5 M arrays compared, the PanAFR performed slightly better in terms of both TRP and FPR. The PanAFR and H3A array out-performed the Omni 2.5 array, despite the similar SNP densities of the three arrays. This likely reveals the importance of enrichment of African-specific common variants for detecting selection signals.

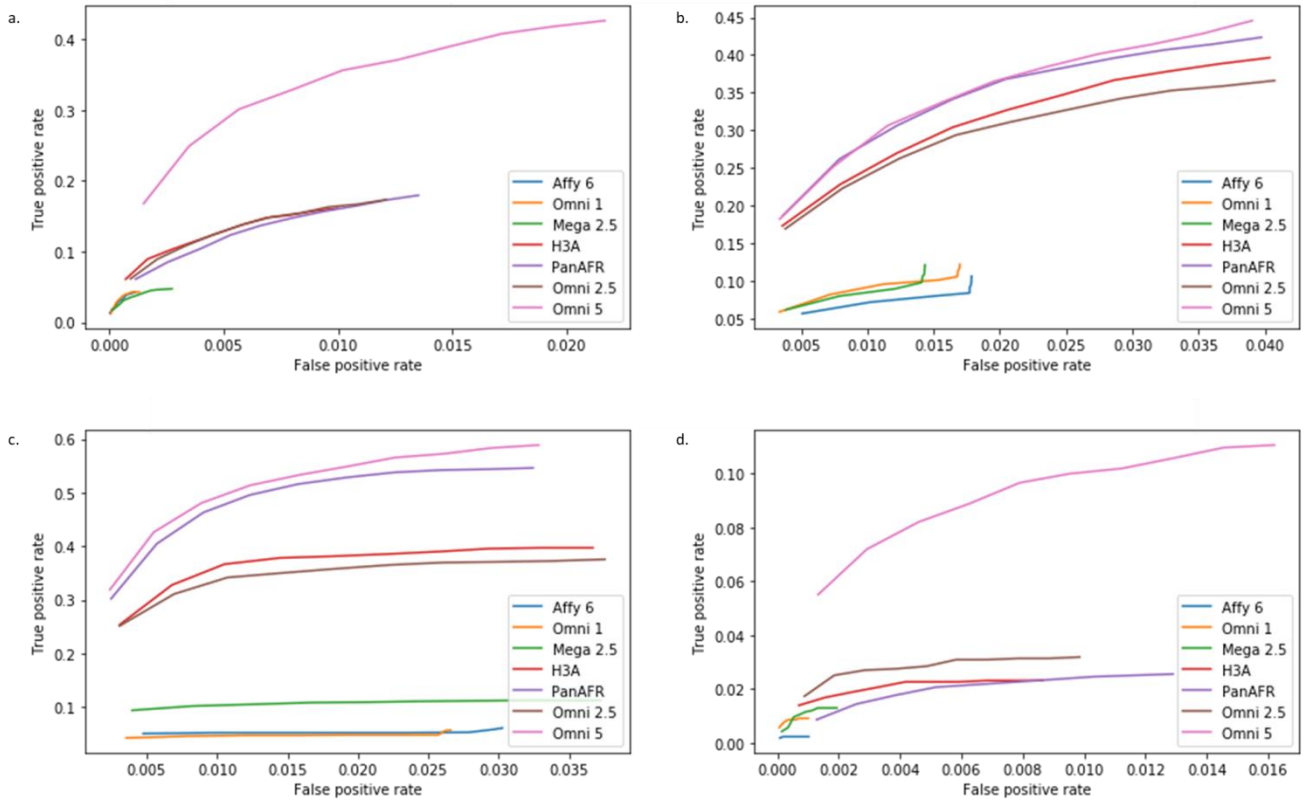

Figure 3.2: The change in true positive rate (TPR) and false positive rate (FPR) of each array over a range of outlier thresholds between 0.5% and 5.0%, in increments of 0.5%, calculated in comparison to the WGS 1% outliers. The thresholds are not explicitly shown because this is a convention for ROC curves, on which these graphs are based. The most SNP-dense arrays tended to produce the highest TPR values, while with most methods, the smallest arrays produced low TPR and FPR values. PanAFR and H3A tended to out-perform Omni 2.5. XP-EHH produced the highest TPR values and Tajima's D the lowest. a. F_{ST}; Ethiopia and Zulu, b. iHS; Zulu, c. XP-EHH; Ethiopia and Zulu, d. Tajima's D; Zulu

The results also indicated that Tajima's D estimates based on chip data might not be very reliable, while the other three methods generated reasonable TPR values which were greater than 30% at an FRP of 0.01.

The gradients of the lines can be used to infer the trade-off between the TPR and FPR, with steeper gradients indicating that the increase in TPR is not counter-balanced by an increase in FPR. The steepest gradients occur between the lowest outlier thresholds, specifically between the 0.05% and 1% thresholds.

3.1.3. The percentage of array outliers which are true positives

To estimate the reliability of selection signals that can be expected with currently available genotyping arrays, the percentage of array 1% outliers which were true positives (TP%) was calculated by comparison to results from WGS data. Since the selection scan results of a smaller array is more likely to contain more FN windows, TPR and FPR measures were clearly impacted by the size of the arrays. Thus, this analysis may provide better estimates of the accuracy of signals produced from array data.

Figure 3.3 shows the TP% values for the Zulu population or Ethiopia and Zulu population pair. Results for the other populations are provided in Appendix 4 and show very similar trends.

The results for F_{ST} showed a clear difference between the trends in the distribution of TP% compared to the trends in TPR seen in the previous two analyses. Although the TPR of the Affy 6 array was found to be very low, at 1.85×10^{-2} (Appendix 2.a.), the TP% values were the highest achieved with any array. Omni 1 also demonstrated a very high TP% compared to other arrays. The highest TP% was 47.54%, for the Affy 6 array, for the pairwise F_{ST} between the Ethiopia and Zulu samples (Figure 3.3.a). The comparison between the TPR and TP% indicate that while the smaller arrays might miss many WGS signals, the outliers identified with these arrays were fairly accurate. This also corresponded with the low FPR values observed for the smallest arrays.

The TP% values produced by the EHH-based methods, however, followed the trends seen with TPR (Figure 3.1.c). As observed with TPR, the PanAFR and H3A arrays produced higher TP% values compared to the similarly sized Omni 2.5 array. The TP% values for the three smaller arrays were very low: for example, only 2.43% of the signals produced by Affy 6 with XP-EHH were TPs.

While the ranking of arrays by TP% was similar among different populations for F_{ST}, iHS and XP-EHH, it differed between populations for the results of Tajima's D (Appendix 4). Moreover, there was almost no relation between array size and TP%. The Omni 1 array was found to produce the highest TP% and clearly out-performed much larger arrays.

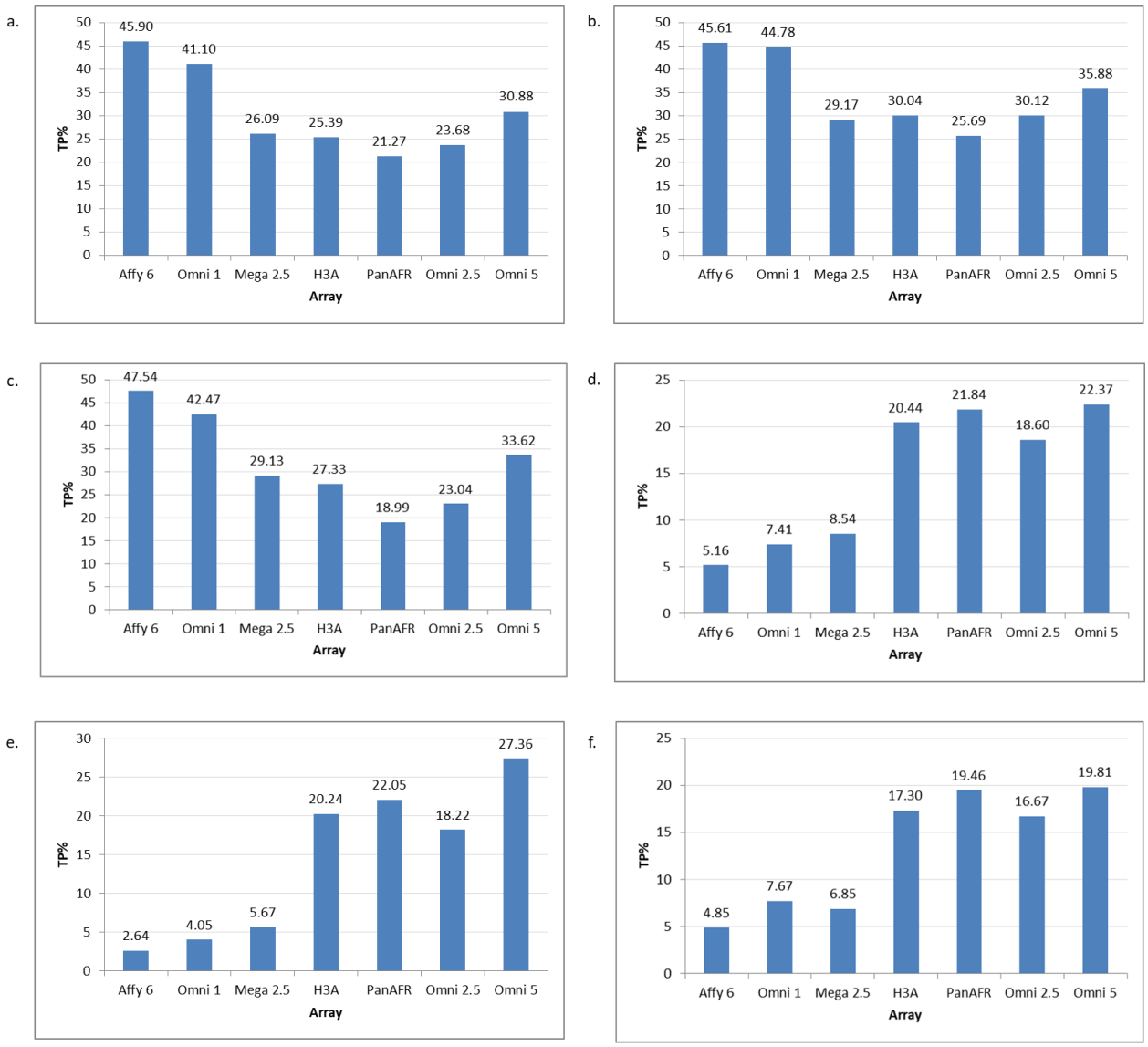

Figure 3.3.1: The percentages of array outliers which were true positives (TP%). With EHH-based methods, larger arrays generally had higher TP% values, while with F_{ST} and Tajima's D, some of the smallest arrays produced the highest values. a. F_{ST}; Baganda and Ethiopia, b. F_{ST}; Baganda and Zulu, c. F_{ST}; Ethiopia and Zulu, d. iHS; Baganda, e. iHS; Ethiopia, f. iHS; Zulu.

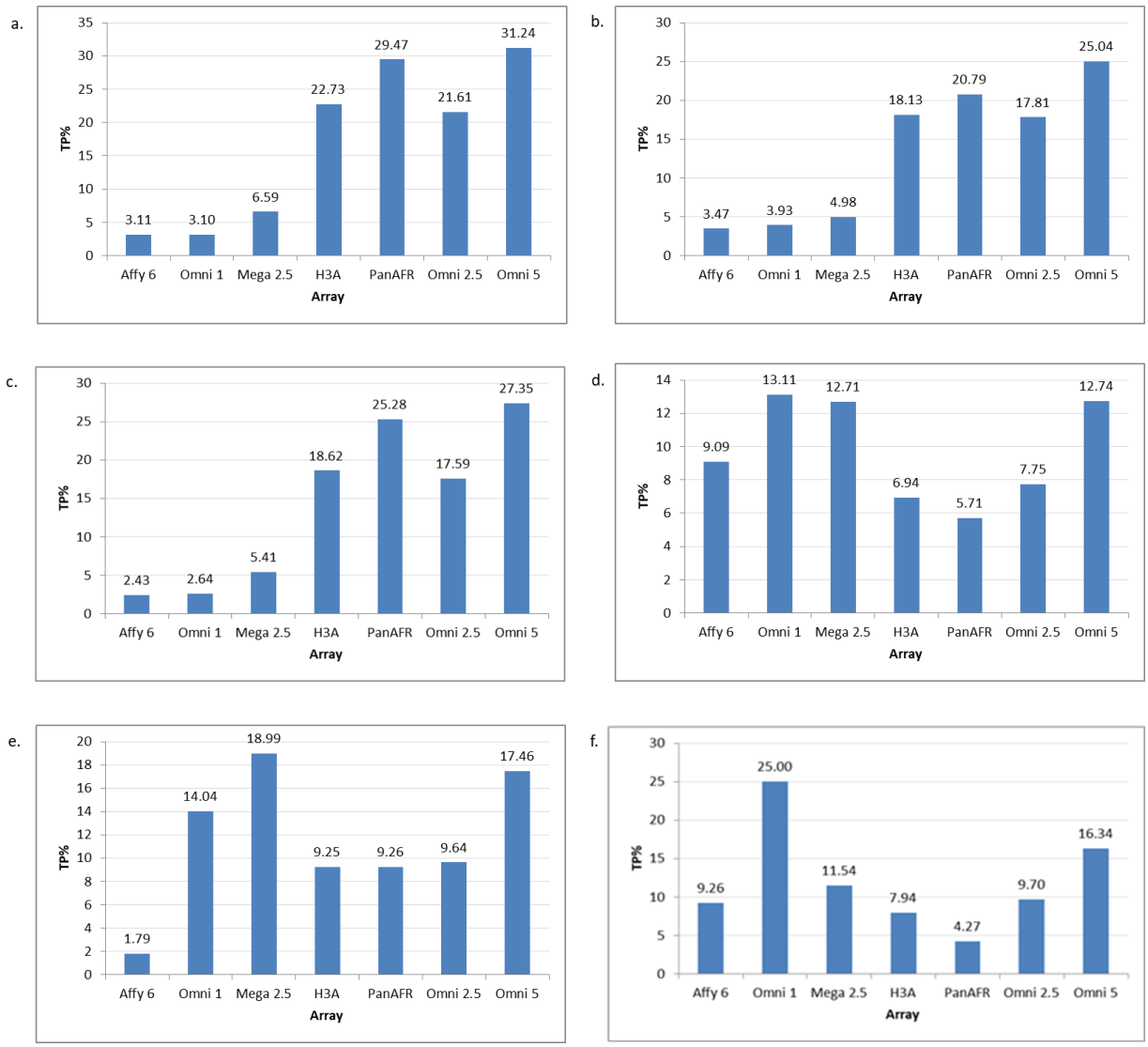

Figure 3.3.2: The percentages of array outliers which were true positives (TP%). With EHH-based methods, larger arrays generally had higher TP% values, while with F_{ST} and Tajima's D, some of the smallest arrays produced the highest values. a. XP-EHH; Baganda and Ethiopia, b. XP-EHH; Baganda and Zulu, c. XP-EHH; Ethiopia and Zulu, d. Tajima's D; Baganda, e. Tajima's D; Ethiopia, f. Tajima's D; Zulu.

Thus, the markers of the Omni 1 array are a good starting point when ascertaining markers for larger arrays. This analysis also revealed that methods based on F_{ST} are the most accurate for detecting signatures of selection, especially when using smaller arrays. The EHH-based methods were found to be accurate only with the larger arrays. Therefore, previously reported results based on 1M or smaller arrays might need to be reinvestigated. The Tajima's D method was again seen to be strongly impacted by the representation of genomic regions by the ascertainment of SNP markers.

3.1.4. The correlation of array and WGS outlier scores

The stark contrast between the trends in TPR and TP% revealed that representations of accuracy can be strongly swayed by the statistics used. To achieve a more direct assessment, the overall correlation between the scores of WGS 1% outliers and the corresponding array score was examined. Anderson-Darling tests performed on subsets of the data indicated that they weren't normally distributed (Appendix 5), so the non-parametric Kendall's tau test was performed (Table 3.1).

Kendall's tau values indicated that the scores for selection statistics of 1% outliers were weakly correlated between array and WGS results. The Kendall's tau values also showed no obvious pattern in array performance across populations and selection statistics. The correlations were visualized by plotting the selection statistic scores of WGS 1% outliers against the scores produced from array data for corresponding windows. Plots for Omni 5 (the largest array) for each selection statistic for the Zulu or Zulu and Ethiopia pair are given in Figure 3.4. Plots for the other populations, arrays and methods are provided in Appendix 6.

Table 3.1: Kendall's Tau values with corresponding p values for the correlation between scores of WGS 1% outliers and scores produced from the array data for the same windows. P values below $5x10^{-6}$ were considered significant to account for multiple testing. Kendall's tau values for the least SNP-dense arrays generally weren't significant and weak correlations were observed for all data subsets. a. F_{ST}, b. iHS, c. XP-EHH, d. Tajima's D.

a.

Population	Array	Kendall's Tau	p value	$p < 5x10^{-6}$
Baganda & Ethiopia	Affy 6	0.30	$2.77x10^{-4}$	
Baganda & Ethiopia	Omni 1	0.28	$4.65x10^{-4}$	
Baganda & Ethiopia	Mega 2.5	0.18	$9.08x10^{-3}$	
Baganda & Ethiopia	H3A	0.26	$1.96x10^{-11}$	*
Baganda & Ethiopia	PanAFR	0.19	$1.37x10^{-8}$	*
Baganda & Ethiopia	Omni 2.5	0.23	$7.48x10^{-11}$	*
Baganda & Ethiopia	Omni 5	0.24	$1.72x10^{-26}$	*
Baganda & Zulu	Affy 6	0.28	$1.78x10^{-3}$	
Baganda & Zulu	Omni 1	0.23	$1.01x10^{-3}$	
Baganda & Zulu	Mega 2.5	0.35	$1.85x10^{-6}$	
Baganda & Zulu	H3A	0.24	$2.25x10^{-11}$	*
Baganda & Zulu	PanAFR	0.15	$4.92x10^{-7}$	*
Baganda & Zulu	Omni 2.5	0.22	$5.24x10^{-11}$	*
Baganda & Zulu	Omni 5	0.20	$1.44x10^{-18}$	*
Ethiopia & Zulu	Affy 6	0.28	$3.18x10^{-4}$	
Ethiopia & Zulu	Omni 1	0.41	$1.88x10^{-8}$	*
Ethiopia & Zulu	Mega 2.5	0.27	$3.07x10^{-5}$	
Ethiopia & Zulu	H3A	0.25	$5.19x10^{-12}$	*
Ethiopia & Zulu	PanAFR	0.16	$1.11x10^{-6}$	
Ethiopia & Zulu	Omni 2.5	0.20	$7.33x10^{-9}$	*
Ethiopia & Zulu	Omni 5	0.23	$7.22x10^{-25}$	*

b.

Population	Array	Kendall's Tau	p value	$p < 5x10^{-6}$
Baganda	Affy 6	0.05	$3.76x10^{-1}$	
Baganda	Omni 1	0.14	$1.31x10^{-3}$	
Baganda	Mega 2.5	0.24	$6.58x10^{-9}$	*
Baganda	H3A	0.11	$7.47x10^{-8}$	*
Baganda	PanAFR	0.15	$7.07x10^{-15}$	*
Baganda	Omni 2.5	0.12	$3.06x10^{-8}$	*
Baganda	Omni 5	0.13	$1.80x10^{-12}$	*
Ethiopia	Affy 6	0.23	$2.43x10^{-3}$	
Ethiopia	Omni 1	0.16	$1.49x10^{-2}$	
Ethiopia	Mega 2.5	0.21	$6.94x10^{-5}$	
Ethiopia	H3A	0.19	$2.73x10^{-14}$	*
Ethiopia	PanAFR	0.21	$3.96x10^{-18}$	*
Ethiopia	Omni 2.5	0.18	$1.45x10^{-11}$	*
Ethiopia	Omni 5	0.23	$7.27x10^{-28}$	*
Zulu	Affy 6	0.07	$2.32x10^{-1}$	
Zulu	Omni 1	0.11	$1.26x10^{-2}$	
Zulu	Mega 2.5	0.27	$1.49x10^{-8}$	*
Zulu	H3A	0.15	$1.62x10^{-12}$	*
Zulu	PanAFR	0.16	$7.06x10^{-14}$	*
Zulu	Omni 2.5	0.14	$1.35x10^{-9}$	*
Zulu	Omni 5	0.17	$2.72x10^{-19}$	*

c.

Population	Array	Kendall's Tau	p value	$p < 5x10^{-6}$
Baganda & Ethiopia	Affy 6	0.12	$9.17x10^{-2}$	
Baganda & Ethiopia	Omni 1	0.18	$1.96x10^{-2}$	
Baganda & Ethiopia	Mega 2.5	0.18	$3.48x10^{-4}$	
Baganda & Ethiopia	H3A	0.27	$9.34x10^{-29}$	*
Baganda & Ethiopia	PanAFR	0.28	$1.48x10^{-43}$	*
Baganda & Ethiopia	Omni 2.5	0.23	$2x10^{-21}$	*
Baganda & Ethiopia	Omni 5	0.28	$1.23x10^{-52}$	*
Baganda & Zulu	Affy 6	0.12	$5.66x10^{-2}$	
Baganda & Zulu	Omni 1	0.23	$4.86x10^{-4}$	
Baganda & Zulu	Mega 2.5	0.15	$3.56x10^{-3}$	
Baganda & Zulu	H3A	0.18	$1.87x10^{-13}$	*
Baganda & Zulu	PanAFR	0.21	$3.84x10^{-21}$	*
Baganda & Zulu	Omni 2.5	0.18	$7.60x10^{-13}$	*
Baganda & Zulu	Omni 5	0.19	$7.11x10^{-22}$	*
Ethiopia & Zulu	Affy 6	0.33	$7.29x10^{-5}$	
Ethiopia & Zulu	Omni 1	0.27	$1.52x10^{-3}$	
Ethiopia & Zulu	Mega 2.5	0.23	$4.66x10^{-5}$	
Ethiopia & Zulu	H3A	0.26	$2.05x10^{-24}$	*
Ethiopia & Zulu	PanAFR	0.28	$7.09x10^{-38}$	*
Ethiopia & Zulu	Omni 2.5	0.24	$1.73x10^{-18}$	*
Ethiopia & Zulu	Omni 5	0.27	$3.16x10^{-41}$	*

d.

Population	Array	Kendall's Tau	p value	$p < 5x10^{-6}$
Baganda	Affy 6	0.24	$3.25x10^{-1}$	
Baganda	Omni 1	0.45	$2.00x10^{-2}$	
Baganda	Mega 2.5	0.06	$6.46x10^{-1}$	
Baganda	H3A	0.13	$1.45x10^{-1}$	
Baganda	PanAFR	0.08	$2.63x10^{-1}$	
Baganda	Omni 2.5	0.22	$4.62x10^{-3}$	
Baganda	Omni 5	0.18	$5.23x10^{-6}$	*
Ethiopia	Affy 6	0.11	$3.41x10^{-10}$	*
Ethiopia	Omni 1	0.25	$1.77x10^{-1}$	
Ethiopia	Mega 2.5	0.30	$2.91x10^{-2}$	
Ethiopia	H3A	0.25	$8.8x10^{-3}$	
Ethiopia	PanAFR	0.15	$3.92x10^{-2}$	
Ethiopia	Omni 2.5	0.18	$3.53x10^{-2}$	
Ethiopia	Omni 5	0.21	$9.19x10^{-9}$	*
Zulu	Affy 6	0.67	$1.23x10^{-2}$	
Zulu	Omni 1	0.25	$9.61x10^{-2}$	
Zulu	Mega 2.5	0.08	$4.58x10^{-1}$	
Zulu	H3A	0.26	$1.99x10^{-3}$	
Zulu	PanAFR	0.05	$4.65x10^{-1}$	
Zulu	Omni 2.5	0.24	$8.66x10^{-4}$	
Zulu	Omni 5	0.20	$1.22x10^{-7}$	*

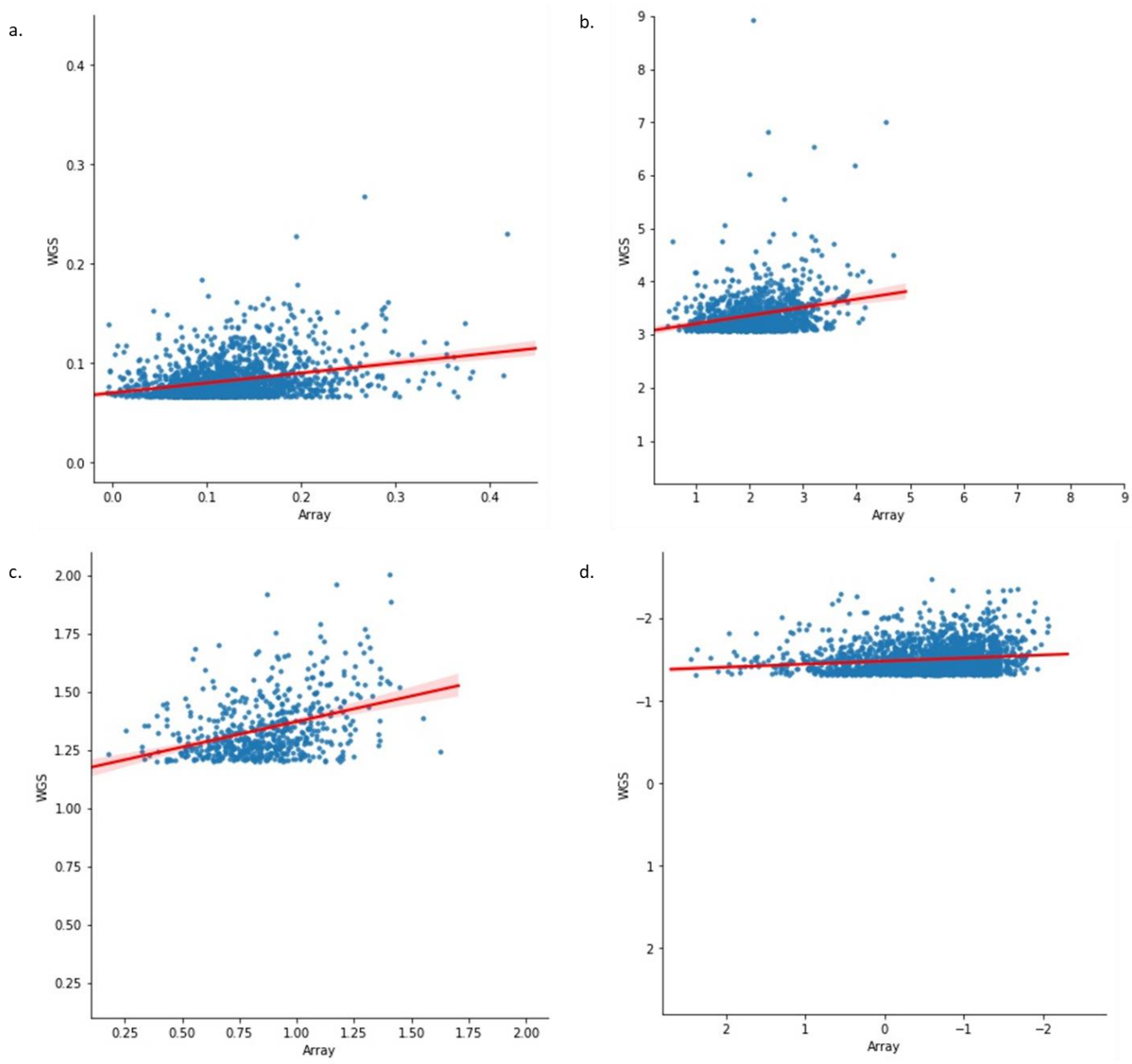

Figure 3.4: The correlation between the selection statistic scores of WGS 1% outlier windows and the scores of corresponding windows from the Omni 5 array results. a. F_{ST}; Ethiopia and Zulu, b. iHS; Zulu, c. XP-EHH; Ethiopia and Zulu, d. Tajima's D; Zulu

The low correlation demonstrated by the Kendall's Tau values was reflected by the weak correlation seen in the scatter plots. WGS outliers generally had higher minimum values than array outliers in all the data subsets. The scatter plots for F_{ST} and Tajima's D showed more variation in the array values, in comparison to the corresponding WGS values.

It should be noted that the scatter plots for the EHH-based arrays were less densely populated, especially for the smaller arrays, because more windows were removed due to the MAF cutoff applied for these methods and because windows with a SNP density <20 were removed from the WGS results.

3.1.5. Concordance between the results of different arrays

The concordance between the selection signals detected using different arrays was compared by representing the number of outlier windows shared by each combination of arrays. The sizes of the intersection of array results were visualized with UpSet plots, which are effectively quantitative Venn diagrams. The plot for the iHS method and Zulu population is shown in Figure 3.5. Results of all other methods and populations are available in Appendix 7.

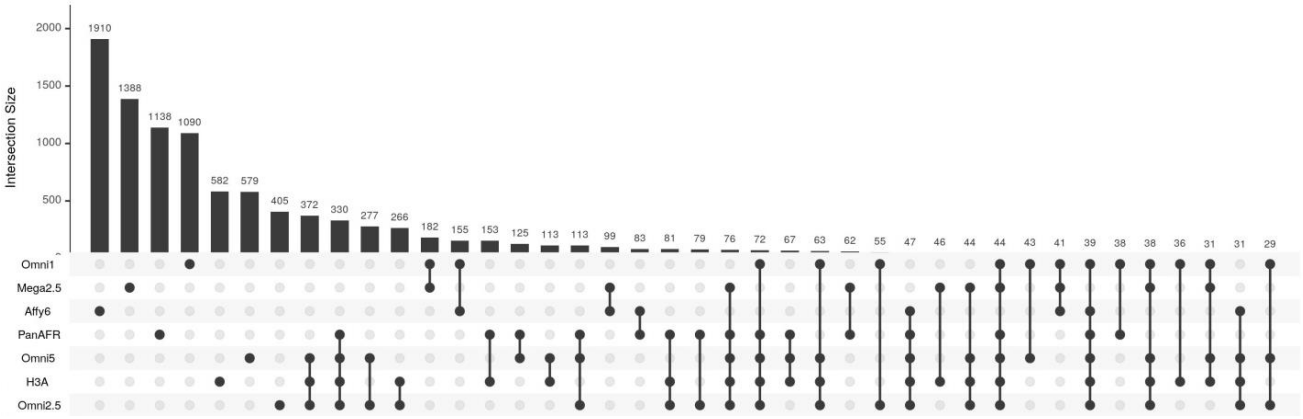

Figure 3.5: The number of outlying windows in the intersection between arrays for the iHS method and Zulu population. For a given combination of arrays, indicated by the dark circles at the bottom of the plot, the number of windows in each intersection is depicted by the bar and the number above it. The majority of signals was identified by a single array, and for this selection statistic the largest intersection consisted of the most SNP-dense arrays.

In accordance with previous results, the largest proportions of windows were each unique to a single array. With the EHH- based methods, the arrays with the fewest markers identified the largest number of unique signals. With F_{ST} and Tajima's D, the largest proportions of windows were unique to the most SNP-dense arrays (Appendix 7). Omni 2.5 and Omni 5 were part of the largest intersection for all the selection statistics and populations, since Omni 2.5 is a subset of Omni 5. PanAFR and H3A tended to be

part of the next largest intersections. The three populations or population pairs showed similar results for each selection scan method.

3.2. The effect of SNP density per window on the detection of signals

To unpack the reason why a given window in array results might be classified either correctly or incorrectly in comparison to WGS results, the distribution of SNP density per window was compared between the accuracy categories. Unmerged and equally-sized windows were considered to enable comparison, and SNP density of array windows was calculated as a percentage of SNPs per WGS window to account for genome-wide variation. The results of the selection scans were taken as input for this analysis, so the criteria applied for the scans were reflected. Most pertinently, only windows with scores for >10 SNPs had been retained, with this threshold increased to >20 SNPs for results of EHH-based tests on WGS data. In EHH-based selection scans, SNPs with MAF<5% had been removed. These criteria were applied to both the WGS and array data. The distributions were visualized with box plots, and plots for H3A (the array with median SNP density) for the Zulu population and Ethiopia and Zulu population pair are provided in Figure 3.6. Plots for the remaining arrays and populations are available in Appendix 8. The accompanying descriptive statistics are provided in Appendix 9. Since many distributions do not appear to be normal, a statistical significance test for the differences in means is not valid. Instead, the Mann-Whitney U statistic was calculated to indicate whether each pair of distributions for each array was significantly different. The U statistics and corresponding p values are given in Appendix 10.

When the distributions were compared between the accuracy categories across the different arrays and populations, no obvious pattern emerged for the F_{ST} statistic (Figure 3.6.a), and Mann-Whitney statistics reached significance for fewer arrays in comparison to other selection statistics.

With the EHH-based methods, across most arrays the FN distributions had the lowest medians, quartile 1 (Q_1) and quartile 3 (Q_3) values, while the FP category had the highest values for these measures (Figure 3.6.b-c and Appendix 8.2.1-8.3.4). The

distributions for the Tajima's D statistic displayed the highest medians, $Q_1$and Q_3 values for the FP category and the lowest for the FN category. Windows with 'positive' classifications tended to have higher SNP densities than windows with 'negative' classifications.

The effect of SNP density was further explored by calculating the percentage of false negatives which were removed from the array data due to a SNP density of less than 10 markers per window. Values for the Zulu population or Ethiopia and Zulu pair are provided in Table 3.2 and tables for the remaining populations are available in Appendix 11.

For all arrays, populations and selection statistics, a very high proportion of false negatives were missed due to low SNP density.

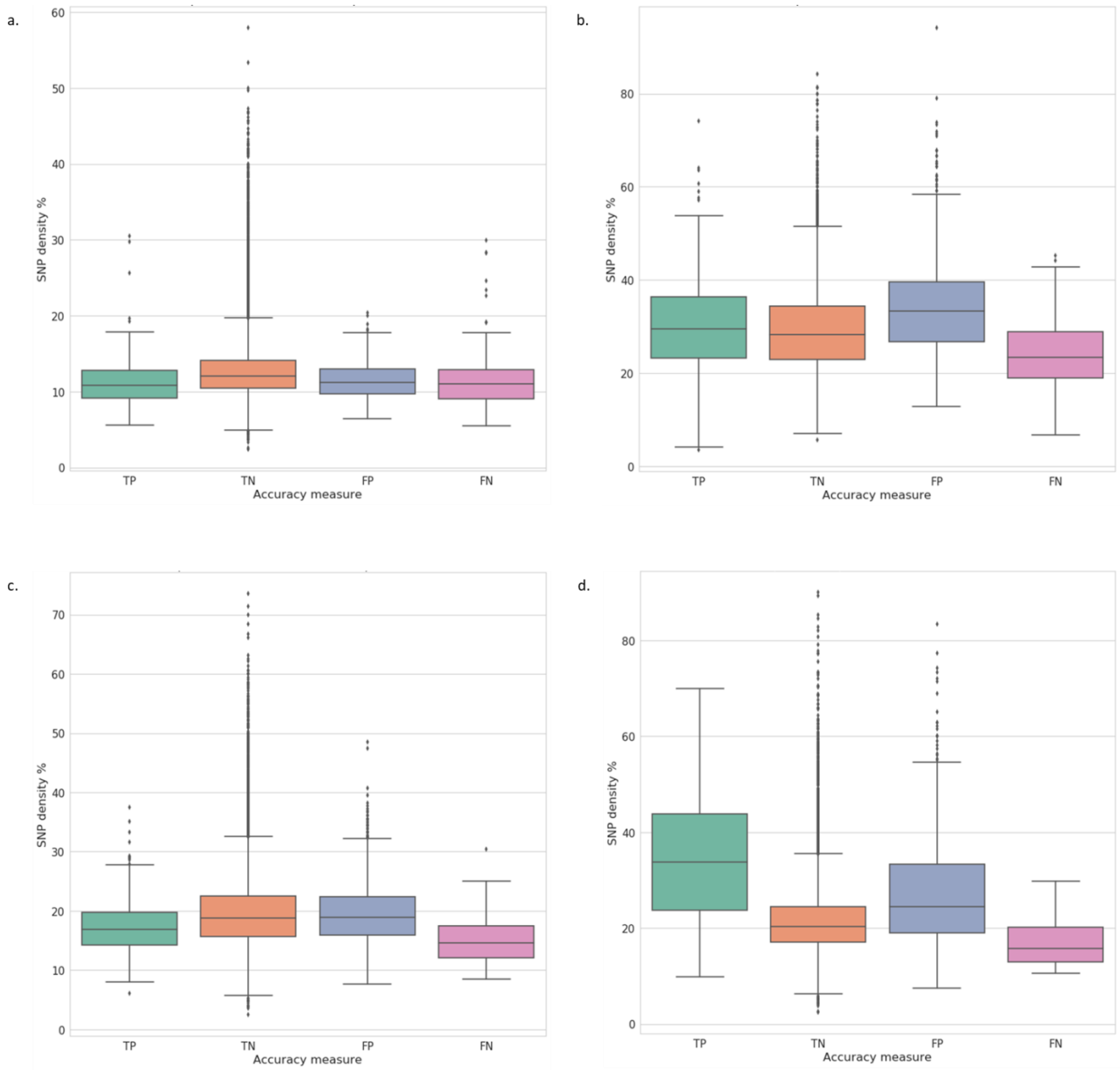

Figure 3.6: The distribution of SNP density per window as a percentage of the number of SNPs in the corresponding WGS window for each accuracy category from the H3A results. The accuracy categories are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Variation in SNP density per category occurred between different arrays, populations and selection statistics, so all plots (Appendix 9) were examined for general patterns. a. F_{ST}; Ethiopia and Zulu; b. iHS; Zulu, c. XP-EHH Ethiopia and Zulu, d. Tajima's D; Zulu.

Table 3.2: Percentages of false negative windows with fewer than 10 SNPs per window. The number of windows with fewer than 10 markers (<10 SNPs) was divided by the total number of false negative windows (FN total) to find the percentage of false negatives with fewer than 10 SNPs per window (%FN <10 SNPs). a. FST, b. iHS, c. XP-EHH, d. Tajima's D.

a.

Population	Array	>10 SNPs	<10 SNPs	FN total	% FN <10 SNPs
Ethiopia & Zulu	Omni 1	47	1511	1558	96.98
Ethiopia & Zulu	Omni 2.5	212	1405	1617	86.89
Ethiopia & Zulu	Omni 5	415	1164	1579	73.72
Ethiopia & Zulu	H3A	176	1406	1582	88.87
Ethiopia & Zulu	Affy 6	44	1511	1555	97.17
Ethiopia & Zulu	PanAFR	260	1409	1669	84.42
Ethiopia & Zulu	Mega 2.5	60	1504	1564	96.16

b.

Population	Array	>10 SNPs	<10 SNPs	FN total	% FN <10 SNPs
Zulu	Omni 1	11	1647	1658	99.34
Zulu	Omni 2.5	212	1401	1613	86.86
Zulu	Omni 5	378	1352	1730	78.15
Zulu	H3A	272	1393	1665	83.66
Zulu	Affy 6	5	1665	1670	99.70
Zulu	PanAFR	272	1332	1604	83.04
Zulu	Mega 2.5	4	1652	1656	99.76

c.

Population	Array	>10 SNPs	<10 SNPs	FN total	% FN <10 SNPs
Ethiopia & Zulu	Omni 1	1	1177	1178	99.92
Ethiopia & Zulu	Omni 2.5	45	831	876	94.86
Ethiopia & Zulu	Omni 5	133	610	743	82.10
Ethiopia & Zulu	H3A	50	800	850	94.12
Ethiopia & Zulu	Affy 6	0	1171	1171	100.00
Ethiopia & Zulu	PanAFR	128	630	758	83.11
Ethiopia & Zulu	Mega 2.5	6	1107	1113	99.46

d.

Population	Array	>10 SNPs	<10 SNPs	FN total	% FN <10 SNPs
Zulu	Omni 1	4	2052	2048	99.81
Zulu	Omni 2.5	32	2015	1983	98.41
Zulu	Omni 5	131	1921	1790	93.18
Zulu	H3A	25	2032	2007	98.77
Zulu	Affy 6	2	2062	2060	99.90
Zulu	PanAFR	57	2037	1980	97.20
Zulu	Mega 2.5	22	2055	2033	98.93

3.3. The concordance between selection scan results from two independent samples of the Southeastern Bantu-speaking population

The second source of ascertainment bias examined was the sample of individuals with which a study represents a population. If population substructure is present, studies of a population from the same geographic region might produce different results. Three selection scans (F_{ST}, iHS and Tajima's D) were performed on both the Bt20 sample and the AGVP Zulu sample, downsized to represent the Omni 5 panel. Outliers within the 0.1%, 0.5% and 1% tails of the empirical score distributions were isolated as selection candidates. The significance of each overlap was determined by finding a p value and representation factor (R factor) value from the hypergeometric distribution. The numbers and percentages of outliers shared by both samples, as well as the R factor values and p values are given in Table 3.3.

Table 3.3: The numbers and percentages of selection candidates at various outlier thresholds shared between the Bt20 and AGVP Zulu results, with Baganda and Ethiopia reference populations for F_{ST}. P values for the significance of the overlap between samples was obtained from the hypergeometric distribution, and these values are significant if the R factor value is above 2. a. F_{ST}; b. iHS; c. Tajima's D

a. F_{ST}

Comparative population	Outlier %	Shared outliers	Total outliers	% Shared outliers	R factor	p value
Baganda	0.1	0	92	0.00	0.0	$p<0.008$
Baganda	0.5	8	465	1.72	3.4	$p<0.003$
Baganda	1.0	22	926	2.38	2.4	$p<2.161\times10^{-4}$
Ethiopia	0.1	2	90	2.22	22.2	$p<0.004$
Ethiopia	0.5	26	452	5.75	11.5	$p<1.322\times10^{-19}$
Ethiopia	1.0	43	892	4.82	4.8	$p<5.495\times10^{-17}$

b. iHS

Outlier %	Shared outliers	Total outliers	% Shared outliers	R factor	p value
0.1	67	220	30.45	304.5	$p<3.684\times10^{-149}$
0.5	369	1137	32.45	64.9	p<0.0
1.0	828	2274	36.41	36.4	p<0.0

c. Tajima's D

Outlier %	Shared outliers	Total outliers	% Shared outliers	R factor	p value
0.1	40	92	43.48	425.3	$p<1.930\times10^{-98}$
0.5	216	453	47.68	95.4	p<0.0
1.0	452	892	50.67	50.7	p<0.0

The low p values and high R factor values indicated that the two datasets are not independent and contain more overlap than expected by chance. However, there were notable differences in overlap observed by signals identified using the four different methods. Selection scans based on F_{ST} produced the lowest levels of overlap between samples, and Tajima's D based estimate showed the highest overlap. Overall, the percentage of overlap among the candidates was observed to be less than in 55% in all cases reflecting largescale differences, at least across certain genomic regions, among the two South African SEB groups.

3.4. The agreement between results of different methods

Since different selection scan methods have been designed to detect selection of various types and at varying time depths (Vitti et al. 2013), it was expected that the four selection scans would identify both common and unique signals.

The extent to which the results of these methods overlap was represented in an UpSet plot for the Zulu population, with Ethiopia as a comparative population for the two-population methods, given in Figure 3.7 and Appendix 7.1-7.4.

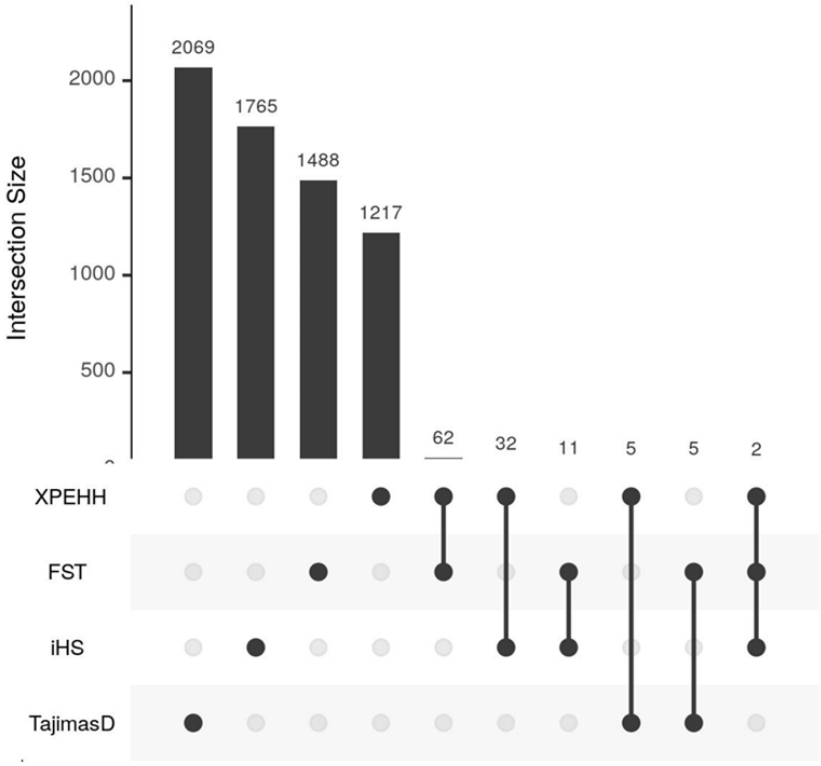

Figure 3.7: The concordance between signals identified by different selection statistics. The heights of the bars indicate the number of WGS selection signals in the intersection of results from various methods. The combinations of selection scan methods are specified by the dark circles below the bars. Disparities are observed between the signals identified by each method.

In spite of using different approaches, the largest intersection was formed between the two inter-population methods, F_{ST} and XP-EHH. The next largest intersection was observed between signals identified by the two EHH-based methods.

3.5. The extent of signal sharing between populations

Various selective pressures have affected the genomic variation of populations living in specific environments, throughout human history. Thus, some signatures of selection are shared by populations which were affected by the same pressure or which are descendants of a common ancestral population, while other selective pressures are only evident in a single population which experienced a unique selective influence.

To compare the overlap between signals identified in the three AGVP populations, the sharing of selection signals was examined using UpSet plots. These plots were based on 1% outliers identified from the WGS data using each selection scan method . Plots for the iHS and XP-EHH methods are given in Figure 3.8, and plots for the remaining selection statistics are available in Appendix 12.

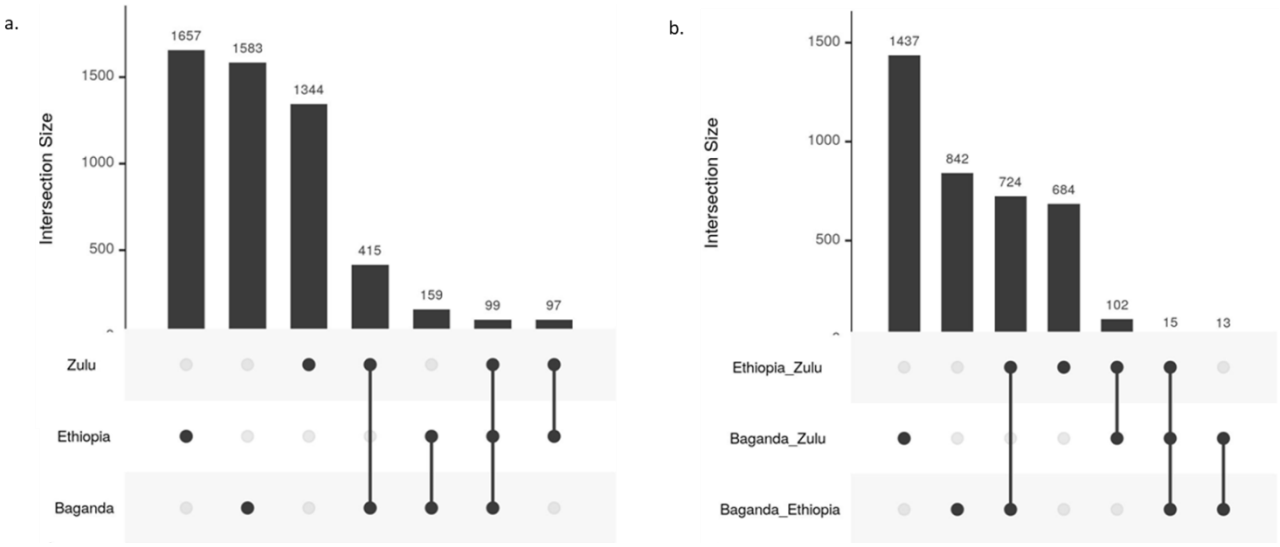

Figure 3.8: The concordance between selection signals observed in the three AGVP populations. The heights of the bars and the numbers above them indicate the number of windows identified as selection signals by the given combination of populations or population pairs, specified by the dark circled below the bars. Most signals were unique to a single population or population pair. a. iHS; b. XP-EHH

As described in the AGVP study, distinctive genetic components of the Ethiopian population distinguish this group from other African populations (Gurdasani et al. 2015). Accordingly, for the single-population selection statistics (Figure 3.8.a), a slightly higher number of unique signals was observed in the Ethiopian sample, while the largest intersection was observed between the Baganda and Zulu populations. For the two-population methods (Figure 3.8.b), the Baganda and Ethiopia population pair contained the most unique signals, while the population pairs including Ethiopia formed the largest intersection, again reflecting the unique ancestry in Ethiopia, compared to the other two populations in this study. While the Baganda and Zulu populations consist of Bantu-speakers, the Ethiopian group contains more distantly related Afro-Asiatic speakers, who are known to have gained genetic diversity from admixture with Eurasian populations.

3.6. Annotation of WGS selection candidates

Outlying WGS windows which were regarded as candidate selected regions were annotated with gene names, gene descriptions and phenotypic descriptions to provide some insight into the reasons why these genes might have been selected for.

3.6.1. Selection candidates which were missed by all arrays and methods

As an example of the effect of ascertainment bias, the strongest selection signals from the WGS results which were not identified from any array by any method were annotated with gene names and listed in Figure 3.9.

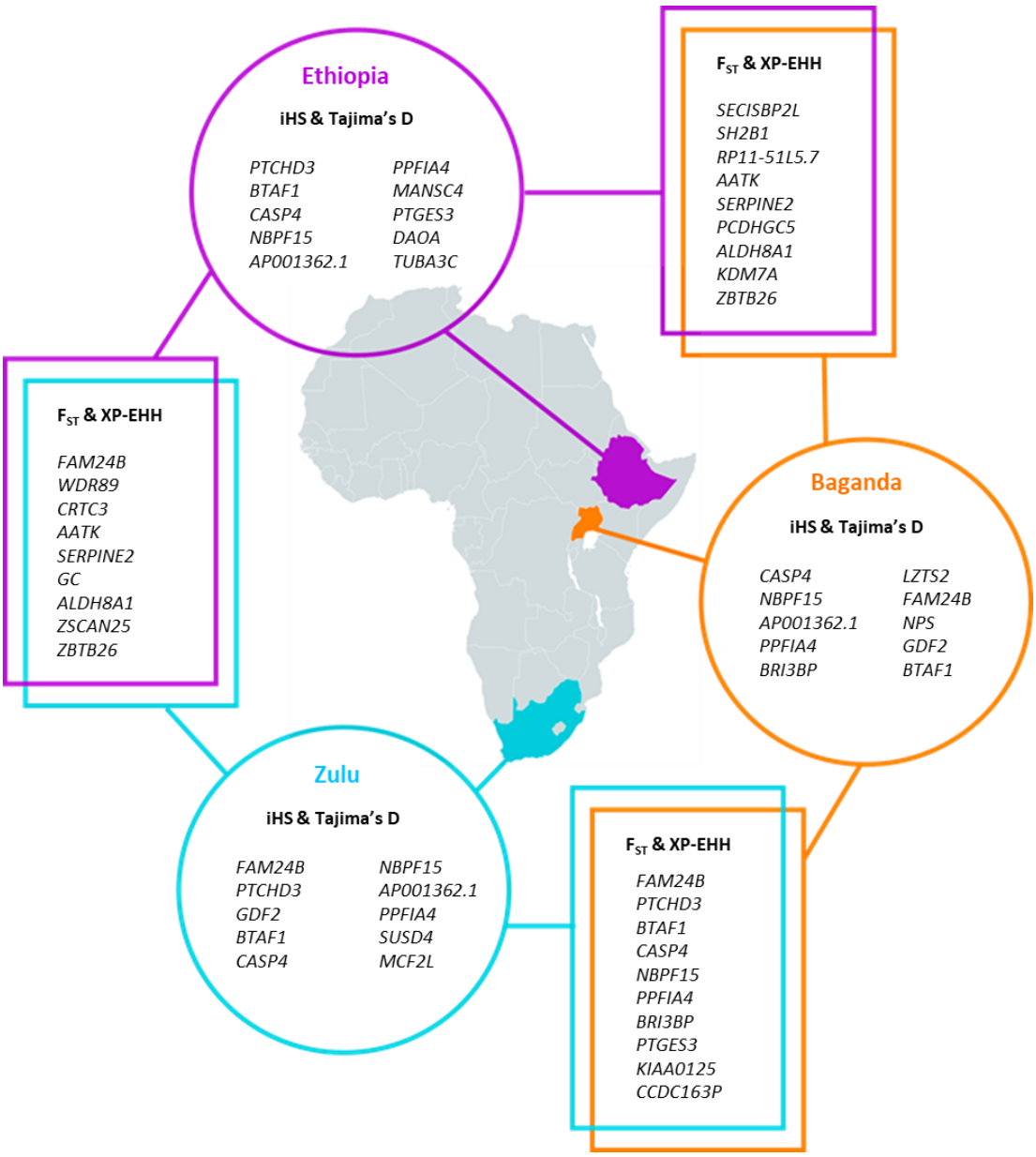

Figure 3.9: Lists of genes which were identified as selection candidates from the WGS 0.01% outliers, but not by any array or selection scan method.

3.6.2. Functional annotations of WGS selection candidates

The top 0.05% WGS selection signals identified in each population using the four selection scan statistics were functionally annotated. The gene descriptions (based on Ensembl GRCh37) for the Zulu population are provided in Table 3.4 and the gene and phenotype descriptions for all populations or population pairs are given in Appendix 13. The Ensembl database does not contain phenotype descriptions for each outlying gene, so Table 3.4 is not the complete list of 0.05% WGS selection signals.

Mutations in many of these genes have been implicated in diseases, indicating their functional significance, which may have led to their selection. Complete lists of selection candidates per selection statistic and population are provided in Appendix 14.

3.6.3. Comparison of selection candidates to previously identified genes

The selection candidates from this study were compared to a modified list of previously selected signals from the dbPSHP database (Choudhury et al. 2017 personal communications). The selection signals identified in the three AGVP populations included a number of novel genes and a full list of these genes is provided in Appendix 15.

Table 3.4: Phenotype descriptions for WGS 0.05% outlying regions detected as signals with both iHS and Tajima's D in the Zulu population.

Gene name	Phenotype description
PER3	Familial advanced sleep-phase syndrome
CAMTA1	Epithelioid hemangioendothelioma
	Non-progressive cerebellar ataxia with intellectual disability
STOX1	Preeclampsia
HK1	Charcot-Marie-Tooth disease type 4G
	Nonspherocytic hemolytic anemia due to hexokinase deficiency
	Neuropathy hereditary, motor and sensory, russe type
	Retinitis pigmentosa 79
KIF1BP	Goldberg-Shprintzen megacolon syndrome
MYPN	Childhood-Onset Slowly Progressive Nemaline Myopathy
	Cap myopathy
	Childhood-onset nemaline myopathy
	Familial isolated dilated & restrictive cardiomyopathy
DGAT2	Autosomal dominant Charcot-Marie-Tooth disease type 2 due to DGAT2 mutation
SERPINH1	Osteogenesis imperfecta type 3/ X
	Preterm premature rupture of the membranes
GPC6	Omodysplasia type 1 (OMOD1)
	Autosomal recessive omodysplasia
TSHR	Hypothyroidism congenital nongoitrous 1
	Athyreosis
	Familial gestational hyperthyroidism
	Familial hyperthyroidism due to mutations in TSH receptor
	Hypothyroidism due to TSH receptor mutations
	Thyroid hypoplasia
	Hyperthyroidism nonautoimmune
MAP2K1	Cardiofaciocutaneous syndrome 3
	Noonan syndrome 1
GFM1	Hepatoencephalopathy due to combined oxidative phosphorylation defect type 1
KIT	human piebaldism
	Acute myeloblastic leukemia with maturation
	Acute myeloid leukemia with abnormal bone marrow eosinophils
	Acute myeloid leukemia with t(8;21)(q22;q22) translocation
	Aleukemic mast cell leukemia
	Bullous diffuse cutaneous mastocytosis
	Classic mast cell leukemia
	Cutaneous mastocytoma
	Gastrointestinal stromal tumor
	Isolated bone marrow mastocytosis
CYP2U1	Hereditary Spastic Paraplegia
	Autosomal recessive spastic paraplegia type 56
HADH	3-Hydroxyacyl-Coenzyme A Dehydrogenase Deficiency
	Hyperinsulinism due to short chain 3-hydroxylacyl-CoA dehydrogenase deficiency
	Hyperinsulemic hypoglycemia familial 4
PDGFRA	Chronic eosinophilic leukemia
	Gastrointestinal stromal tumor
	Myeloid/lymphoid neoplasm associated with PDGFRA rearrangement
	Precursor B-cell acute lymphoblastic leukemia
	Primary hypereosinophilic syndrome
	Hypereosinophilic syndrome idiopathic
CA2	Osteoporosis autosomal recessive type 3
	Osteopetrosis with renal tubular acidosis

4. Discussion

4.1. Comparison of low coverage whole genome sequence data and array data

Low coverage (<5x) WGS data is a common data source in genome-wide studies. When identifying low-frequency variants, it is more economic and effective to sequence a larger sample with low coverage than to perform high-coverage (>30x) sequencing on fewer individuals (Nielsen et al. 2011). In this study, low coverage WGS data was considered as a 'gold standard' against which to compare array data. However, error rates of low coverage WGS data has been reported to be as high as 15%- significantly greater than for array data (Goodwin et al. 2016). These error rates can be caused by incorrect base-calling and misalignment. Low-coverage sequencing also carries a high likelihood of sampling an allele from only one of two haploid chromosomes. Under-representation of low-frequency alleles is expected to reduce population differentiation measures such as FST and XP-EHH, while positively skewing Tajima's D estimates, causing signatures of selection to be overlooked. Missing rare alleles might increase EHH and create false positives in iHS scans. The availability of high coverage WGS data for African populations in the future will allow for a more accurate comparison to be performed.

4.2. The effect of pooling Ethiopian sub-populations

Four Ethiopian populations were pooled to increase the sample size to 72. This created a sample size similar to the Baganda (n=100) and Zulu (n=100) samples. It is debatable whether more power is gained by pooling closely related populations or keeping them separate, especially since there is significant divergence between these Ethiopian populations. iHS loses power when sample size is below 40 chromosomes, while XP-EHH requires at least 20 chromosomes before power is lost (Pickrell et al. 2009), so the pooling may have been necessary. However, creating a structured sample might have affected selection statistics. It has been demonstrated through simulation studies that population structure decreases the power of iHS and XP-EHH to detect selection signals (Vatsiou et al. 2016). If a haplotype is selected in one sub-population but not in the second, the signal is likely to be masked in the pooled population. iHS would lose power

due to the reduced frequency of the selected haplotype in the population affected by selection, while XP-EHH would lose power due to the increased frequency of the selected haplotype in the population which wasn't affected by selection. Conversely, population structure can increase the proportion of rare alleles, resulting in negative values of Tajima's D and mimicking a signature of selection (Cadzow et al. 2014). If subpopulations adapted to different conditions, signatures similar to balancing selection would present in the pooled population. Population structure might decrease the differences between comparative populations in FST scans, masking selection signals.

4.3. The accuracy with which various genotyping arrays represent whole genome sequence variation in selection studies

This study investigated a number of potential confounders in selection studies, and the primary focus was the accuracy with which genotyping arrays represent WGS data.

Available genotyping arrays have improved greatly over the last decade. The first mode of progress involved the increase in array size from one million to around five million SNPs. The primary 1M arrays were designed to capture common variants across the genome, and due to the availability of Eurocentric data, these arrays were enriched for European variants. WGS data from largescale sequencing projects such as the KGP together with technological innovations led to the addition of SNPs to constitute larger arrays, such as Omni 2.5. These larger arrays provided increased coverage of potentially functional variants as well as some representation of variants that are unique to particular populations or common only in some continental groups. The second phase of improvement in array design was the ascertainment of continent-specific variation and haplotypes, leading to Asian- and African-based arrays. For example, the PanAFR and H3A arrays were created with the purpose of increasing coverage of variants in African populations. This design contrasted with the ascertainment of the Omni 5 array, which was purposed to represent SNPs of intermediate frequencies (MAF>0.05). These differences in array design resulted in SNP panels which differentially represent certain variant classes and genomic regions.

Due to the accessibility of array data, selection studies have been conducted on a variety of SNP panels. This presented the need to investigate the level of variation in signals produced by different arrays. While some strong signals have been reproduced across studies, disparities in results have been reported for separate studies of closely related samples. This study aimed to investigate whether such differences could result from unequal coverage of genomic regions between arrays. Although high coverage population-scale WGS data is an ideal standard, the low coverage WGS data which was available for this study was a sufficiently unbiased reference to compare the results from a variety of arrays.

Array data contains much fewer SNPs than WGS data, so many genomic windows are represented with poor coverage, or are not represented at all. In this study, windows with a SNP density below ten were discarded as a QC measure, since selection statistics based on these could not be interpreted with confidence. Therefore, many 10kb regions which might have been selection candidates could not be included in the array-based analysis. Selection signals were identified as windows in the outlying 1% of the selection statistic score distribution, and so the number of windows in 1% tail is proportional to the total number of windows in the dataset. Since smaller arrays contained fewer windows, the number of outliers from the results of each array necessarily differed and could not be compared directly.

The array and WGS outliers were first compared in terms of TPR and FPR measures. Arrays necessarily missed 'true' signals, as identified by WGS outliers, since fewer windows were represented by an array, and therefore the number of windows in the 99th percentile of the score distribution was smaller. Although an array may perform well in terms of the percentage of outliers which are 'true' signals, the necessity of applying an outlier threshold to identify extreme scores in comparison to the genome-wide distribution results in the dropout of true signals. Consequently, discrepancies will inevitably arise when identifying selection candidates with the outlier approach using genotyping array data rather than WGS data.

The four methods employed in this study identify different features of a selection signal, including allele frequencies. EHH-based methods only include common SNPs, and therefore, results from WGS and array data would be differentially affected only if the

two data sources contained significant differences in the content of common SNPs. Conversely, the F_{ST} method relies on both common and rare alleles and is susceptible to genomic coverage of intermediate-frequency and rare SNPs in array data. Of the four methods, Tajima's D is most dependent on rare variants to detect selection signals, and is expected to be most strongly affected by the under-representation of rare alleles in array data.

In accordance with these expectations, the XP-EHH method produced the highest overall TPR values, while the results of Tajima's D contained the lowest values. This observation is in agreement with previous reports which show that haplotype-based methods are less affected by ascertainment bias than SNP-based methods (Granka et al. 2014). This analysis also suggests that statistics which rely on derived allele frequencies such as Tajima's D are the least reliable when WGS data is not available (Chen et al. 2010; Ferrer-Admetlla 2014).

The TPR and TP% values, used to assess array accuracy, were both considerably low, indicating that many selection signals identified from WGS data were not detectable in any of the array-based data. Comparison of arrays by TPR and FPR values further revealed that the size of the arrays strongly influenced the accuracy of results. The high TP% values of the smallest arrays was surprising. It could be that the markers of smaller arrays were ascertained to more accurately represent known functional regions, which are more likely to be selected. Thus, while many windows dropped out due to sparse sampling of markers, the small number of windows in the top 1% of the selection statistic score distribution contained a large proportion of true signals.

Arrays generally over-represent common SNPs, so are expected to bias Tajima's D upwards, and lower F_{ST} values. Lower SNP density might have led to truncation of extended haplotypes. Thus, lower SNP density would cause selection signals to be missed, and as the proportion of TPs is reduced, the proportion of FPs increases. Additionally, this study removed windows with fewer than 10 SNPs per 10kb window, since estimates based on few SNPs are highly variable. This caused the dropout of many windows, and would have had a greater impact on smaller arrays.

As expected, Omni 5 which contained both the largest number of markers and more rare alleles was found to be most accurate. The second-best performing array, PanAFR, was designed to capture variants with MAF>2% in the Yoruba population. The H3A array shares approximately 1.5M SNPs with the Omni 2.5 array (Choudhury 2017 personal communications), but presumably due to greater representation of African variants, H3A produced more accurate results than Omni 2.5.

The superior performance of African-based arrays compared to a more Eurocentric array of similar size emphasizes the importance of using arrays ascertained from African discovery samples when studying African populations. Due to the high levels of diversity both within and between African populations (Sherman et al. 2018), and the localization of selective events, it is likely that many signatures of selection remain to be discovered on this continent (Campbell et al. 2014; Gomez et al. 2014). The PanAfr array was ascertained from the Yoruba population, while the H3A array ascertainment populations included the Yoruba as well as African populations from both the AGVP and KGP. The H3A array was expected to more accurately represent the samples in this study, so the slightly better performance of the PanAfr array in these selection scans was surprising.

The choice of outlier threshold is subjective, since it is unknown what proportion of a genome is influenced by positive selection. Examining the trade-off between TPR and FPR can assist in finding the threshold which yields the most accurate results. The TPR and FPR were compared at different outlier thresholds for each array, by plotting the two metrics against each other and observing the gradients of the curves. The steepest gradients occurred between the 0.05% and 0.1% outlier thresholds, indicating that the choice of a 1% outlier threshold, rather than 0.05% will yield the greatest increase in accuracy. The largest array produced positive gradients between all outlier thresholds, while for the smaller arrays with XP-EHH, increasing the outlier threshold yielded almost no increase in TPR. Overall, the gradients decreased at the highest outlier thresholds, and this is expected, since the most extreme values of the statistics provide the strongest evidence of positive selection. A lower outlier threshold is preferable, since high false positive rates could invalidate the discovery of true signals.

Secondly, the percentage of array outliers which were true positives was considered. With the EHH-based methods, the TP% measure ranked arrays into roughly the same

order as the TPR metric. The TP% was correlated with the number of markers per array, with the exception of the African based arrays, which out-ranked Omni 2.5. While the TPR values suggested that EHH-based methods are most robust to SNP density, the TP% values for the smallest arrays were very low. Surprisingly, with F_{ST}, the TP% values of the smallest arrays were relatively high. The comparison between the TPR and TP% for each array indicated that while the smallest arrays missed many WGS signals, the outliers identified with these arrays were fairly accurate. F_{ST} also produced the lowest FPR values across all arrays. Overall, the percentages of outliers which were true positives were fairly low.

As a third analysis of the effect of ascertainment bias, the correlation between scores of WGS 1% outliers and the scores of corresponding windows from array results was considered. These results suggest that in addition to the rank of the signals, the ascertainment of bias could also lead to variation in signal strengths. Therefore, the strengths of signals inferred from array data do not provide a completely reliable approximation of selection strength in WGS data.

The fourth assessment of ascertainment bias involved comparing the overlap between signals from different arrays. The largest proportions of windows were each unique to a single array. This indicates that the differences in coverage across the genome strongly impact the results of selection studies. The low levels of overlap between the outliers of different arrays may contribute to the discordance in results of different studies which has been reported (Voight et al. 2006; Hernandez et al. 2007; Oleksyk et al. 2010; Fagny et al 2014).

For all assessment of ascertainment bias, results for a given selection statistic and array tended to be similar for the three different populations or population pairs. This increased confidence that the observed trends were properties of a given selection scan method or array.

TP signals were regarded as 'true' in relation to the WGS results, but signatures of selection aren't a falsifiable hypothesis, so it isn’t certain that genomic regions displaying evidence of selection have truly been selected for (Pavlidis et al. 2012). Multiple confounders such as demographic effects and purifying selection are known to create

patterns of variation similar to those of a selective sweep (Granka et al. 2012), and these alternative causes can't be ruled out with the selection scans which were used.

Selection scans using array data excluded signals consisting of functional regions which may be important selection candidates. This was demonstrated by Figure 3.9. Some of these candidates which were detected in all three populations from WGS data but not from any arrays are highlighted below.

CASP4 encodes the protein Caspase 4, which belongs to the cysteine-aspartic acid protease (caspase) family. This protein family is involved in cellular apoptosis (Martinon & Tschopp 2007). *NBPF15* belongs to the neuroblastoma breakpoint family (NBPF), which contain tandemly repeated copies of DUF1220 protein domains. These duplications have been associated with multiple developmental and neurogenetic diseases (Vandepoele et al. 2005). *APO01362*.1 is part of the APO gene group encodes apolipoproteins, which bind lipids to enable transportation (Saito et al. 2004). *PPFIA4* belongs to the liprin-alpha gene family and encodes the protein Liprin-alpha-4. It is involved in pathways including Neurotransmitter Release Cycle and Transmission across Chemical Synapses (Ko et al. 2003).

4.4. The effect of SNP density per window on the detection of selection signals

The analysis of the correlation between accuracy classification (especially FN and FP) and the SNP density of a region suggest that SNP coverage in array data strongly affects the detection of selection signals. For both the EHH-based methods, across all arrays, FN windows tended to have lower SNP density, while FP windows contained a higher number of markers.

Low SNP density in or around a selection signal may have led to the truncation of an extended haplotype, causing a signal to be missed. Similarly, homozygous regions represented by more markers could have appeared proportionately longer, leading to false positive signals around regions of high homozygous SNP density. Qanbari et al.

(2010) studied EHH in cattle genomes and likewise found that regions with higher SNP density were more prone to be identified as a core region by EHH-based methods.

With the Tajima's D statistic, 'positive' windows tended to have higher SNP densities than 'negative' windows, suggesting that the method is less likely to classify a window as a selection signal if it has a low SNP density. Results of the F_{ST} scans showed no clear pattern in SNP density per accuracy category.

Arrays have generally been designed to represent the genome in association studies, so SNPs are sampled more sparsely in regions with high LD. Pickrell et al. (2009) analyzed data produced by the Human Genome Diversity Panel containing 650,000 common SNPs and found that the array has fewer markers in selected regions than the genome-wide average. This element of ascertainment bias may contribute to false negatives in selection studies.

Given the high FN rates observed, there are many genes which were not detected in selection scans on array data. This was exemplified by the lists of the strongest WGS selection candidates which were not detected from array data, in Figure 3.9. The functionality of these genes is unlikely to be the reason why they did not produce extreme scores of selection statistics. Rather, the FN regions were annotated to provide concrete examples of selection candidates which would be overlooked as an example of ascertainment bias and low SNP density.

4.5. The concordance between results from two independent samples of the Southeastern Bantu-Speaking group

In addition to the AGVP Zulu data analyzed in this study, the availability of genotype data from May et al. (2013), enabled a comparison of the signals of selection scan in two different SEB samples and comment on homogeneity of signals in the two groups. While, there is overall high genetic proximity between the AGVP Zulu and Bt20 samples, both of which are South Eastern Bantu speakers, the unexpectedly low concordance between signals from these two samples suggests some form of differentiation between these groups.

These differences could result from the sampling locations of the two groups: individuals sampled by the AGVP study all self-identified as Zulu and were sampled from Durban, whereas the Bt20 cohort was sampled from Soweto, which is a more diverse region, including ethnicities such Zulu, Sotho, Venda and Xhosa. Soweto has attracted migrant workers from around the country since the gold era in South Africa and continued urbanization has caused the Sowetan population to assimilate individuals from distinct geographic regions.

ADMIXTURE analysis performed in previous studies which included these samples has shown that in spite of the similarity of the Bt20 (May et al. 2013) and AGVP (Gurdasani et al. 2015) individuals to other southeastern bantu-speaking individuals, the Bt20 shows considerably higher KhoeSan gene flow in comparison to the Zulu (Choudhury et al. 2017). This difference in the level of KhoeSan gene flow can therefore be hypothesized to be the major source of the observed differences in selection signals between the two populations. These results suggest that there might be sufficient genetic differentiation within the SEB and that signatures of selection might vary between subgroups of SEB depending on geography, and the extent of KhoeSan gene flow.

The KhoeSan have a higher proportion of rare SNPs than any other population, and under-representation of this diversity by arrays is likely to cause selection signals inherited from the Khoesan to be missed. Additionally, selected regions inherited from recent admixture events are more likely to occur on longer haplotypes which have not yet been disrupted by recombination, so the timing of admixture between the Khoesan and different SEB sub-groups could influence the selection signals detected. Differential admixture with the Khoesan might contribute to varying selection signals between SEB samples. For example, the Xhosa have a greater proportion of Khoesan ancestry than other SEB language groups such as the Zulu (Petersen et al. 2013). XP-EHH and F_{ST}, which rely on population differentiation to detect selection, and admixture could either increase or decrease the differentiation between comparative populations. The introduction of differentially selected haplotypes could weaken iHS signals, and the introduction of rare alleles from the more diverse KhoeSan population could create false positives in Tajima's D scans. The effect of admixture on selection scans could be investigated further with a method which is robust to admixture, such as pcadapt (Luu et

al. 2017). This method relies on principal components analysis (PCA) to detect selection candidates as outliers with respect to population structure.

4.6. The amount of signal sharing between populations

Overall, low overlap was observed between the selection signals from the three AGVP populations. The genome-wide average F_{ST} estimates in Table 1.1 indicate that the Baganda and Zulu are more closely related to each other than to the Ethiopian sample. The AGVP pooled Ethiopian sample includes Afro-asiatic speakers with Eurasian ancestry that distinguishes them from the Bantu-speaking populations (Gurdasani et al. 2015). The divergence between Afro-Asiatic speakers and other sub-Saharan African populations occurred approximately 50kya (Shriner et al. 2014), and selection at this time depth is expected to be detected by all the statistics used in this study. The migration of Bantu-speakers into sub-Saharan Africa occurred 2-3kya (Patin et al. 2017), and more recent sweeps which occurred since the divergence between the Zulu and Baganda populations are most likely to be detected by EHH-based methods. The two Bantu-speaking populations studied are also distinguished by differential gene flow, with the Baganda population including ancestral contributions from East-African, which the Zulu lack. In contrast the Zulu population has experienced a low level of KhoeSan gene flow which is absent in the both the East African populations. Therefore, varying degree of non-Bantu speaker (predominantly East African) ancestry in the three groups could explain the marginal overlap between signals detected in the three populations.

4.7. The agreement between results of different selection scan methods

The low concordance observed between the signals detected using various methods could reflect the different time depths at which these methods detect selection. For example Tajima's D detects the most ancient sweeps (250kya), while F_{ST} identifies more recent sweeps (50kya). It has been demonstrated that no single method is able to detect both starting and nearly completed selective sweeps (Vatsiou et al. 2016b) and even methods designed to detect the same type of sweeps have been found to produce dissimilar results (Akey et al. 2009). Additionally, various methods could produce

different false positives and false negatives, decreasing the concordance between results (Schrider & Kern 2017).

4.8. Annotation of some common selection candidates in three African populations

Selection scans produce long lists of candidate regions, and once these are identified the challenge becomes to decipher the selective advantage which a genomic variant has provided in a given environment. This study identified both canonical selection candidates and genes for which evolutionary histories have yet to be reconstructed. The following genes are well-known selection signals, with convincing adaptive evolutionary narratives.

The selection candidates identified in this study included one of the first discovered instances of positive selection in humans. The *HBB* gene was identified as an extended haplotype in the Ethiopian population. Haldane (1949) first observed that the geographic distribution of the sickle haemoglobin (*HbS*) allele mirrored the range of malaria infection in Africa, and hypothesized that the allele was driven to high frequency by the resistance it confers to malaria infection in the heterozygous carrier state Haldane (1949).

Signatures of selection have been discovered around other variants which protect against malaria, including the T188G mutation in platelet glycoprotein 4 (*CD36*) (Pain et al. 2001). This study observed differential selection on *CD36* between the Baganda and Ethiopia populations. CD36 is expressed on the surface of platelets where it binds multiple ligands including lipoproteins and erythrocytes infected with *Plasmodium falciparum* (Ferrer-Admetlla et al. 2014). An extended haplotype has been found to overlap this gene exclusively in African populations (Sabeti et al. 2006; Patin et al. 2017).

The *LARGE* gene, which encodes an N-acetylglucosaminyltransferase gene, occurred within an extended haplotype in all three populations studied. Selection on *LARGE* has been attributed to the differential susceptibility to the Lassa virus. However, this virus is endemic to West Africa (Andersen et al. 2011), so the extended haplotypes observed could have been inherited from the recent West African ancestors.

Infectious diseases are known to be one of the strongest selective pressures, and many selection candidates have immune-related functions. The *HLA* complex is another classic selection candidate (Sabeti et al. 2006; Patin et al. 2017) identified in the Baganda and Ethiopian populations. This gene complex encodes the major histocompatibility complex (*MHC*) proteins which are essential to adaptive immunity. *HLA* genes consist of some of the most rapidly evolving genomic sequences and it is believed that this variability developed as a mechanism of 'herd immunity' to prevent entire populations from being wiped out by a single strain of pathogen (Wills & Green 1995).

The Zulu population displayed evidence of selection on the *IL34* gene, which encodes interleukin-34, a cytokine that promotes the differentiation and viability of monocytes and macrophages, and so has an important role in immunity. This was also the strongest signal found by Ferrer-Admetlla et al. (2014) in the Yoruba. Another signature in the Baganda and Zulu occurred around *TLR5*. This gene encodes Toll-like receptor 5, which recognizes bacterial flagellin and plays a role in innate immunity. It has also been identified as a strong selection signal in the Yoruba (Fagny et al. 2014).

A number of genes associated with various cancers were identified as selection candidates. These included *NUP214*, *MSH2*, *EPCAM*, *KIT* and *PDGFRA*. Enrichment for somatic mutations associated with cancer has previously been identified among selection candidates (Schrider & Kern 2017). It has been estimated that 5.4% of genes associated with cancer display evidence of positive selection (Weghorn & Sunyaev 2017). This is believed to be the result of genomic conflict, in which alternative forces of selection favour opposing mechanisms. Tumour suppression would be selected to enable survival, while on the cellular level, cancer progression might be favoured to prevent apoptosis during spermatogenesis. It has also been suggested that the extended human lifespan has presented a selective pressure which prevents cancer progression (Nunney & Muir 2015).

Adaptation to high altitude (>2,500 meters) is another classic selective pressure. Physiological phenotypes including hemoglobin levels are known to have high heritability and greatly improve fitness at high altitudes. Previously identified signatures of selection were found in the *VAV3*, *ARNT2* and *THRB* genes, which have been

associated with hemoglobin levels in Ethiopia (Scheinfeldt et al. 2012). *ARNT2* forms part of the HIF-1 pathway, which is also under selective pressure in the elevated Tibetan and Andean environments.

UV radiation levels have driven local adaptation around the world. Light skin pigmentation enables vitamin D production in regions with less UV exposure, while dark pigmentation provides protection from over-exposure. *HERC2* has previously been associated with variation in pigmentation among African populations (Crawford et al. 2017) and displayed signatures of selection in the Baganda and Ethiopian populations. The rs4932620 (T) allele is associated with dark skin pigmentation and is most frequent in Ethiopian individuals with an increased proportion of Nilo-Saharan ancestry.

Many selection candidates identified in this study have been associated with neurological and neuro-motor disorders. These include *DYRK1A*, *SYT1*, *GRIK2*, *GRID2*, *ITPR1*, *NRXN3*, *CAMTA1*, and *CNTNAP2*. The association indicates that these genes have essential functions and may have been involved in cognitive and neuro-motor developments in the human lineage. For example, *FOXP2* displayed evidence of positive selection in the Ethiopian and Zulu populations. This gene is involved in synapse regulation, maturation and density, and multiple lines of evidence have linked *FOXP2* to human-specific speech development (Sousa et al. 2017).

A number of taste receptor genes, including *TAS2R14*, *TAS2R31* and *TAS2R1* displayed signatures of selection. These genes are expressed in the taste receptor cells in the tongue and are associated with bitter taste perception (Behrens et al. 2004). The *TAS2R* family of genes has been linked to differences in dietary preferences, and bitter taste perception may have evolved as a preventative measure to ingesting toxic plants (Wooding et al. 2004).

While convincing stories of adaptive evolution can be drawn from some selection candidates, in other cases it is difficult to infer the cause of the selection signal. A large region spanning multiple genes may show evidence of selection, with no indication of which gene is selected for, while other candidate regions contain no genes or regulatory regions. Complete understanding of the role of adaptive evolution will require thorough

investigation of the candidates identified, by detecting a molecular or phenotypic change associated with genetic variants in the selected region.

However, the validation of selection candidates by their functional importance has been criticized. Pavlidis et al. (2012) demonstrated the limitation of this approach by simulating neutral population histories, performing selection scans on these neutral genomes, and functionally annotating the false positive selection signals. They showed that plausible narratives of adaptive evolution can be created for neutral regions, and argued that functionality doesn't validate a selection candidate.

4.9. Study limitations

Ideally, the significance of outliers should be assessed by comparing evidence from both empirical genome-wide distributions and theoretical model distributions produced by simulations. Thus, this study could be strengthened by identifying selection candidates via additional comparison to a neutral model. Commonly used arrays were compared to WGS data in order to demonstrate the accuracy of results from real arrays. However, this analysis could be supplemented by simulating random SNP panels of similar size to the real arrays, performing selection scans on these datasets, and assessing whether the purposeful ascertainment of array markers produce superior results to arbitrary sets of markers. When DNA samples are genotyped by real arrays, many markers fall out due to technical reason. Thus, performing this study on data produced by real genotyping arrays would provide more accurate results. Additionally, imputation could be performed on the array data to compare the accuracy of selection scan results to those of the original array data. For the XP-EHH method, determining ancestral and derived alleles by comparison to an outgroup would have revealed which population was affected by selection. The selection scan statistics used are under- powered to detect balancing selection and soft sweeps. A three- way F_{ST} statistic such as LSBL would have indicated which population was affected by selective pressures. This study was performed using low coverage WGS data, but rare variants can often only be called by deep sequencing. Only high-coverage sequence data will allow confidence in studies of highly variable loci, some of which are critical to human evolution. This study performed four of the most commonly used selection scans, but there are many other statistical

tests which could be compared. Windows of 10kb containing fewer than 10 SNPs were discarded, but alternatively, a larger window size could have been used to ensure a sufficient number of SNPs per window (Pickrell et al. 2009). Additional widely used arrays such as the Affymetrix Human origins array could have been included. The annotation of selection signals could be improved by performing a GO analysis using software such as DAVID, TOPPGENE or GOWINDA, rather than only providing a list of GO terms. Finally, larger sample sizes would have strengthened confidence in the results.

4.10. Future prospects

Breakthroughs continue to be made in in this field and the central challenges of selection studies are being addressed. New statistical tests are continuously being developed and compared to well-established methods (Vatsiou et al. 2016b). Simulation programs (Ewing & Hermisson 2010) are accounting for more complex non-equilibrium demographic scenarios which can confound detection of selection signals. Methods based on approximate Bayesian computation (Bazin et al. 2010) and deep learning (Sheehan & Song 2016) have been developed to simultaneously infer demography and selection, possibly overcoming one of the greatest confounders in selection studies. Incorporating background selection into models is one limitation which remains to be addressed, and increasing availability of computational resources will aid in this pursuit. Attention has shifted to polygenic selection in which multiple genes, pathways and groups of genes with related functions each contribute small effects (Stephan 2016). Genomic annotation and functional validation of selection candidates will continue to supplement the hypotheses generated by selection scans. Greater sequencing efforts will continue to shed light on African evolutionary histories. Many ethnolinguistic groups in Africa have not yet been studied, and given the genetic and phenotypic diversity in Africa, there is much left to discover about selection on the continent (Campbell & Tishkoff 2010). The rapidly falling cost of WGS data will remedy ascertainment bias and enable more reliable results in selection studies.

5. Conclusions

The current study is to the knowledge the first investigation on the transferability of signals between different genotyping arrays as well as between genotyping arrays and WGS data. Comparisons of the results from selection scans performed on both WGS and genotyping array panels have indicated that ascertainment bias has a strong impact on the results of selection studies. Selection scans performed on widely used high-density genotyping arrays produced many false negatives and false positives, and array size and ascertainment bias appeared to be the strongest determinants of signal accuracy. The two African-based arrays provided an exception to this trend, performing slightly better than a more Eurocentric array with a comparable number of markers. This highlights the importance of using an array which was ascertained from populations with similar ancestry to the genotyped population.

This study suggests that when full sequence data is not available, use of EHH-based selection statistics could provide more reliable estimates of selection, compared to the SFS-based approach, as the latter is more strongly affected by low marker density. Limited overlap was observed between the results of different methods and arrays, indicating these variables could contribute to the previously reported disparities between the selection candidates of studies on single populations.

Additionally, selection scans performed on two independent SEB subgroups showed significant differences in the top signals between the two samples. The differences between the Bt20 and AGVP Zulu samples could indicate that the SEB group is genetically complex. Alternatively, the reduced overlap of selection signals could reflect the limitation of comparing down-sampled low coverage sequence data to real genotyping array data.

6. References

1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. *Nature, 526*(7571), p.68.

Akey, J.M., 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, 19(5), pp.711-722.

Albrechtsen, A., Nielsen, F.C. and Nielsen, R., 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, 27(11), pp.2534-2547.

Andersen, K.G., Shylakhter, I., Tabrizi, S., Grossman, S.R., Happi, C.T. and Sabeti, P.C., 2012. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. Philosophical Transactions of the Royal Society B: *Biological Sciences*, 367(1590), p.868.

Atkinson, Q.D., Gray, R.D. and Drummond, A.J., 2009. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1655), pp.367-373.

Bank, C., Ewing, G.B., Ferrer-Admettla, A., Foll, M. and Jensen, J.D., 2014. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*, 30(12), pp.540-546.

Bazin, E., Dawson, K.J. and Beaumont, M.A., 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, 185(2), pp.587-602.

Behrens, M., Brockhoff, A., Kuhn, C., Bufe, B., Winnig, M. and Meyerhof, W., 2004. The human taste receptor hTAS2R14 responds to a variety of different bitter compounds. *Biochemical and Biophysical Research Communications*, 319(2), pp.479-485.

Beltrame, M.H., Rubel, M.A. and Tishkoff, S.A., 2016. Inferences of African evolutionary history from genomic data. *Current Opinion in Genetics & Development*, 41, pp.159-166.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., Herráez, D.L. and Brutsaert, T., 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genetics*, 6(9), p.e1001116.

Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H. and Stephan, W., 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2), pp.783-796.

Cadzow, M., Boocock, J., Nguyen, H.T., Wilcox, P., Merriman, T.R. and Black, M.A., 2014. A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, 5, p.293.

Campbell, M.C. and Tishkoff, S.A., 2010. The evolution of human genetic and phenotypic variation in Africa. *Current Biology*, 20(4), pp.166-173.

Campbell, M.C., Hirbo, J.B., Townsend, J.P. and Tishkoff, S.A., 2014. The peopling of the African continent and the diaspora into the new world. *Current Opinion in Genetics & Development*, 29, pp.120-132.

Chen, H., Patterson, N. and Reich, D., 2010. Population differentiation as a test for selective sweeps. *Genome Research*, 20(3), pp.393-402.

Chimusa, E.R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H. and Ramesar, R., 2015. A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. *PLoS Genetics*, 11(3), p.e1005052.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E.R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M.J. and Joubert, F., 2017. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, 8(1), pp.2062-2074.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. and Nielsen, R., 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15(11), pp.1496-1502.

Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., Garrison, E., Xue, Y. and Tyler-Smith, C., 2014. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, 15(6), p.R88.

Conway, J.R., Lex, A. and Gehlenborg, N., 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18), pp.2938-2940.

Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. and Pritchard, J.K., 2009. The role of geography in human adaptation. *PLoS Genetics*, 5(6), p.e1000500.

Crawford, N.G., Kelly, D.E., Hansen, M.E., Beltrame, M.H., Fan, S., Bowman, S.L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y. and Pfeifer, S.P., 2017. Loci associated with skin pigmentation identified in African populations. *Science*, 358(6365), pp. 887-901

Delaneau, O., Coulonges, C. and Zagury, J.F., 2008. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9(1), pp.540-554.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. and McVean, G., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156-2158.

Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M. and Excoffier, L., 2013. Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution*, 30(7), pp.1544-1558.

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E. and Blum, M.G., 2015. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. *Molecular Biology and Evolution*, 33(4), pp. 1082-1093.

Eichstaedt, C.A., Antão, T., Pagani, L., Cardona, A., Kivisild, T. and Mormina, M., 2014. The Andean adaptive toolkit to counteract high altitude maladaptation: genome-wide and phenotypic analysis of the Collas. *PLoS One*, 9(3), p.e93314.

Enard, D., Messer, P.W. and Petrov, D.A., 2014. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6), pp.885-895.

Ewing, G. and Hermisson, J., 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16), pp.2064-2065.

Fagny, M., Patin, E., Enard, D., Barreiro, L.B., Quintana-Murci, L. and Laval, G., 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Molecular Biology and Evolution*, 31(7), pp.1850-1868.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R., 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), pp.1275-1291.

Goodwin, S., McPherson, J.D. and McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), p.333.

Gomez, F., Hirbo, J. and Tishkoff, S.A., 2014. Genetic variation and adaptation in Africa: implications for human evolution and disease. *Cold Spring Harbor Perspectives in Biology*, 6(7), p.a008524.

Granka, J.M., Henn, B.M., Gignoux, C.R., Kidd, J.M., Bustamante, C.D. and Feldman, M.W., 2012. Limited evidence for classic selective sweeps in African populations. *Genetics*, pp.1049-1064.

Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A., 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10), p.1031.

Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H. and Cabili, M., 2013. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4), pp.703-713.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A. and Ritchie, G.R., 2015. The African genome variation project shapes medical genetics in Africa. *Nature*, 517(7534), pp.327-901.

Haasl, R.J. and Payseur, B.A., 2016. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, 25(1), pp.5-23.

Haldane, J.S., 1949. The rate of mutation of human genes. *Hereditas*, 35(S1), pp.267-273.

Henn, B.M., Steele, T.E. and Weaver, T.D., 2018. Clarifying distinct models of modern human origins in Africa. Current opinion in genetics & development, 53, pp.148-156.

Hernandez, R.D., Williamson, S.H. and Bustamante, C.D., 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Molecular Biology and Evolution*, 24(8), pp.1792-1800.

Holsinger, K.E. and Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, 10(9), pp.639-650.

International HapMap Consortium, 2003. The international HapMap project. *Nature*, 426(6968), p.789.

Jarvis, J.P., Scheinfeldt, L.B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.M., Beggs, W., Hoffman, G. and Mezey, J., 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genetics*, 8(4), p.e1002641.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W. and Akey, J.M., 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8), pp.980-989.

Ko, J., Na, M., Kim, S., Lee, J.R. and Kim, E., 2003. Interaction of the ERC family of RIM-binding proteins with the liprin-α family of multidomain proteins. *Journal of Biological Chemistry*, 278(43), pp.42377-42385.

Lachance, J. and Tishkoff, S.A., 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, 35(9), pp.780-786.

Lane, A.B., Soodyall, H., Arndt, S., Ratshikhopha, M.E., Jonker, E., Freeman, C., Young, L., Morar, B. and Toffie, L., 2002. Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. *American Journal of Physical Anthropology*, 119(2), pp.175-185.

Lewontin, R.C. and Krakauer, J., 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1), pp.175-195.

Li, M.J., Wang, L.Y., Xia, Z., Wong, M.P., Sham, P.C. and Wang, J., 2014. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Research*, 42(D1), pp.D910-D916.

Liu, X., Ong, R.T.H., Pillai, E.N., Elzein, A.M., Small, K.S., Clark, T.G., Kwiatkowski, D.P. and Teo, Y.Y., 2013. Detecting and characterizing genomic signatures of positive selection in global populations. *The American Journal of Human Genetics*, 92(6), pp.866-881.

Lund, J. (2005). *Statistical significance of overlap of two groups of genes*. [online] Nemates.org. Available at: http://nemates.org/MA/progs/overlap_stats.html [Accessed 15 Dec. 2018].

Luu, K., Bazin, E. and Blum, M.G., 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular ecology resources*, *17*(1), pp.67-77.

MacLeod, I.M., Hayes, B.J. and Goddard, M.E., 2014. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics*, 198(4), pp.1671-1684.

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P. and Aurelle, D., 2016. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology*, 25(1), pp.170-184.

Martinon, F. and Tschopp, J., 2007. Inflammatory caspases and inflammasomes: master switches of inflammation. *Cell death and differentiation*, 14(1), p.10.

Matise, T.C., Chen, F., Chen, W., Francisco, M., Hansen, M., He, C., Hyland, F.C., Kennedy, G.C., Kong, X., Murray, S.S. and Ziegle, J.S., 2007. A second-generation combined linkage–physical map of the human genome. *Genome Research*, 17(12), pp.1783-1786.

May, A., Hazelhurst, S., Li, Y., Norris, S.A., Govind, N., Tikly, M., Hon, C., Johnson, K.J., Hartmann, N., Staedtler, F. and Ramsay, M., 2013. Genetic diversity in black South Africans from Soweto. *BMC Genomics*, 14(1), p.644.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. and Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11), pp.1566-1575.

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), p.443.

Nunney, L. and Muir, B., 2015. Peto's paradox and the hallmarks of cancer: constructing an evolutionary framework for understanding the incidence of cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1673), p.20150161.

Oleksyk, T.K., Smith, M.W. and O'Brien, S.J., 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1537), pp.185-205.

Orr, H.A., 2009. Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, 10(8), pp.531-539.

Pain, A., Urban, B.C., Kai, O., Casals-Pascual, C., Shafi, J., Marsh, K. and Roberts, D.J., 2001. A non-sense mutation in Cd36 gene is associated with protection from severe malaria. *The Lancet*, 357(9267), pp.1502-1503.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., Froment, A. and Heyer, E., 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, 356(6337), pp.543-546.

Pavlidis, P., Jensen, J.D., Stephan, W. and Stamatakis, A., 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, 29(10), pp.3237-3248.

Petersen, D.C., Libiger, O., Tindall, E.A., Hardie, R.A., Hannick, L.I., Glashoff, R.H., Mukerji, M., Fernandez, P., Haacke, W., Schork, N.J. and Hayes, V.M., 2013. Complex patterns of genomic admixture within southern Africa. PLoS genetics, 9(3), p.e1003309.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. and Pritchard, J.K., 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome research*, *19*(5), pp.826-837.

Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B. and Reich, D., 2014. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111(7), pp.2632-2637.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), pp.559-575.

Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R. and Simianer, H., 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Animal genetics*, 41(4), pp.346-356.

Quinlan, A.R. and Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841-842.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S., 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834), pp.199-204.

Ramírez-Soriano, A. and Nielsen, R., 2009. Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics*, 181(2), pp.701-710.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. and Ackerman, H.C., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909), p.832.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S., 2006. Positive natural selection in the human lineage. *Science*, 312(5780), pp.1614-1620.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. and Schaffner, S.F., 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164), p.913.

Saito, H., Lund-Katz, S. and Phillips, M.C., 2004. Contributions of domain structure and lipid interaction to the functionality of exchangeable human apolipoproteins. *Progress in lipid research*, 43(4), pp.350-380.

Scheinfeldt, L.B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., Lambert, C., Jarvis, J.P., Abate, D., Belay, G. and Tishkoff, S.A., 2012. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biology*, 13(1), R1.

Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G. and Soodyall, H., 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, 338(6105), pp.374-379.

Schrider, D.R. and Kern, A.D., 2017. Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, 34(8), pp.1863-1877.

Segurel, L. and Bon, C., 2017. On the evolution of lactase persistence in humans. Annual Review of Genomics and Human Genetics, 18, pp.297-319.

Sethupathy, P. and Hannenhalli, S., 2008. A tutorial of the poisson random field model in population genetics. *Advances in Bioinformatics*, 2008, http://dx.doi.org/10.1155/2008/257864.

Sheehan, S. and Song, Y.S., 2016. Deep learning for population genetic inference. *PLoS Computational Biology*, 12(3), p.e1004845.

Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E. and Levin, A.M., 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nature genetics, 51(1), p.30-35.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), pp.308-311.

Shriner, D., Tekola-Ayele, F., Adeyemo, A. and Rotimi, C.N., 2014. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. Scientific reports, 4, p.6055.

Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M. and Jones, K.W., 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human genomics*, *1*(4), p.274.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G. and Bardou, P., 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1), pp.W589-W598.

Sousa, A.M., Meyer, K.A., Santpere, G., Gulden, F.O. and Sestan, N., 2017. Evolution of the human nervous system function, structure, and development. *Cell*, 170(2), pp.226-247.

Stephan, W., 2016. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, 25(1), pp.79-88.

Szpiech, Z.A. and Hernandez, R.D., 2014. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, 31(10), pp.2824-2827.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), pp.585-595.

Teshima, K.M., Coop, G. and Przeworski, M., 2006. How reliable are empirical genomic scans for selective sweeps?. *Genome Research*, *16*(6), pp.702-712.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324: 1035–1044.

Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. and Van Roy, F., 2005. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Molecular Biology and Evolution*, 22(11), pp.2265-2274.

Vatsiou, A.I., Bazin, E. and Gaggiotti, O.E., 2016. Changes in selective pressures associated with human population expansion may explain metabolic and immune related pathways enriched for signatures of positive selection. *BMC Genomics*, 17(1), https://doi.org/10.1186/s12864-016-2783-2.

Vatsiou, A.I., Bazin, E. and Gaggiotti, O.E., 2016. Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular Ecology*, 25(1), pp.89-103.

Veeramah, K.R., Wegmann, D., Woerner, A., Mendez, F.L., Watkins, J.C., Destro-Bisol, G., Soodyall, H., Louie, L. and Hammer, M.F., 2011. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Molecular Biology and Evolution*, *29*(2), pp.617-630.

Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K., 2006. A map of recent positive selection in the human genome. *PLoS Biology*, 4(3), p.e72.

Weghorn, D. and Sunyaev, S., 2017. Bayesian inference of negative and positive selection in human cancers. *Nature Genetics*, 49(12), p.1785.

Weir, B.S. and Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), pp.1358-1370.

Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. and Hill, W.G., 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, 15(11), pp.1468-1476.

Wills, C. and Green, D.R., 1995. A genetic herd-immunity model for the maintenance of MHC polymorphism. *Immunological Reviews*, 143(1), pp.263-292.

Wooding, S., Kim, U.K., Bamshad, M.J., Larsen, J., Jorde, L.B. and Drayna, D., 2004. Natural selection and molecular evolution in PTC, a bitter-taste receptor gene. *The American Journal of Human Genetics*, 74(4), pp.637-646.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G. and Gil, L., 2017. Ensembl 2018. *Nucleic acids research*, 46(D1), pp.D754-D761.

Zhai W, Nielsen R, Slatkin M. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution*. 2009 Feb 1;26(2):273-83.

7. Description of Appendices

Appendix 1.1.

Ethics certificate for this MSc project

Appendix 1.2.

Ethics certificate for the Bt20 data

Appendix 1.3.

Ethics certificate for the AGVP data

Appendix 2

The number of genomic windows in each accuracy category and the true positive rates (TRP) and false positive rates (FPR) which were calculated from these counts. Windows in a certain percentage (outlier %) of a selection statistic score distribution were considered ‘postive’ signals and were compared to the whole genome sequence 1% outliers, which were considered as ‘true’ signals. Each window was classified as either true positive (TP), true negative (TN), false positive, or false negative (FN). A table of values is provided for each selection statistic and population or pair of populations.

Appendix 2.1.

a. F_{ST}; Baganda and Ethiopia
b. F_{ST}; Baganda and Zulu

Appendix 2.2.

a. F_{ST}; Ethiopia and Zulu
b. iHS; Baganda

Appendix 2.3.

a. iHS; Ethiopia
b. iHS; Zulu

Appendix 2.4.

a. XP-EHH; Baganda and Ethiopia
b. XP-EHH; Baganda and Zulu

Appendix 2.5.

a. XP-EHH; Ethiopia and Zulu
b. Tajima's D; Baganda

Appendix 2.6.

a. Tajima's D; Ethiopia
b. Tajima's D; Zulu

Appendix 3

The positions of WGS 1% outliers in the array score distributions, examined by plotting the true positive rate (TPR) against the false positive rate (FPR) for each array over an outlier threshold ranging from 0.5% to 5%

a. F_{ST}; Baganda and Ethiopia
b. F_{ST}; Baganda and Zulu
c. iHS; Baganda
d. iHS; Ethiopia
e. XP-EHH; Baganda and Ethiopia
f. XP-EHH; Baganda and Zulu
g. Tajima's D; Baganda
h. Tajima's D; Ethiopia

Appendix 4

The percentage of array outliers which were true positives (TP%) per array.

a. F_{ST}; Baganda and Ethiopia
b. F_{ST}; Baganda and Zulu
c. iHS; Baganda
d. iHS; Ethiopia

e. XP-EHH; Baganda and Ethiopia
f. XP-EHH; Baganda and Zulu
g. Tajima's D; Baganda
h. Tajima's D; Ethiopia

Appendix 5

Normality tests for the WGS score distribution, including only windows represented by the Omni 5 array. Normality was determined with the Anderson-Darling statistic (Statistic) with critical values at various significance levels (Significance level: Critical values).

Appendix 6

The correlation of the selection statistic scores of WGS 1% outlier windows with the scores of corresponding windows from the array results.

Appendix 6.1.

a. F_{ST}; Baganda and Ethiopia; Affy 6
b. F_{ST}; Baganda and Ethiopia; Omni1
c. F_{ST}; Baganda and Ethiopia; Mega 2.5
d. F_{ST}; Baganda and Ethiopia; H3A
e. F_{ST}; Baganda and Ethiopia; PanAFR
f. F_{ST}; Baganda and Ethiopia; Omni 2.5
g. F_{ST}; Baganda and Ethiopia; Omni5

Appendix 6.2.

a. F_{ST}; Baganda and Zulu; Affy 6
b. F_{ST}; Baganda and Zulu; Omni1
c. F_{ST}; Baganda and Zulu; Mega 2.5
d. F_{ST}; Baganda and Zulu; H3A
e. F_{ST}; Baganda and Zulu; PanAFR
f. F_{ST}; Baganda and Zulu; Omni 2.5

g. F_{ST}; Baganda and Zulu; Omni5

Appendix 6.3.

a. F_{ST}; Ethiopia and Zulu; Affy 6
b. F_{ST}; Ethiopia and Zulu; Omni1
c. F_{ST}; Ethiopia and Zulu; Mega 2.5
d. F_{ST}; Ethiopia and Zulu; H3A
e. F_{ST}; Ethiopia and Zulu; PanAFR
f. F_{ST}; Ethiopia and Zulu; Omni 2.5

Appendix 6.4.

a. iHS; Baganda; Affy 6
b. iHS; Baganda; Omni1
c. iHS; Baganda; Mega 2.5
d. iHS; Baganda; H3A
e. iHS; Baganda; PanAFR
f. iHS; Baganda; Omni 2.5
g. iHS; Baganda; Omni5

Appendix 6.5.

a. iHS; Ethiopia; Affy 6
b. iHS; Ethiopia; Omni1
c. iHS; Ethiopia; Mega 2.5
d. iHS; Ethiopia; H3A
e. iHS; Ethiopia; PanAFR
f. iHS; Ethiopia; Omni 2.5
g. iHS; Ethiopia; Omni5

Appendix 6.6.

a. iHS; Zulu; Affy 6
b. iHS; Zulu; Omni1

c. iHS; Zulu; Mega 2.5
d. iHS; Zulu; H3A
e. iHS; Zulu; PanAFR
f. iHS; Zulu; Omni 2.5

Appendix 6.7.

a. XP-EHH; Baganda and Ethiopia; Affy 6
b. XP-EHH; Baganda and Ethiopia; Omni1
c. XP-EHH; Baganda and Ethiopia; Mega 2.5
d. XP-EHH; Baganda and Ethiopia; H3A
e. XP-EHH; Baganda and Ethiopia; PanAFR
f. XP-EHH; Baganda and Ethiopia; Omni 2.5
g. XP-EHH; Baganda and Ethiopia; Omni5

Appendix 6.8.

a. XP-EHH; Baganda and Zulu; Affy 6
b. XP-EHH; Baganda and Zulu; Omni1
c. XP-EHH; Baganda and Zulu; Mega 2.5
d. XP-EHH; Baganda and Zulu; H3A
e. XP-EHH; Baganda and Zulu; PanAFR
f. XP-EHH; Baganda and Zulu; Omni 2.5
g. XP-EHH; Baganda and Zulu; Omni5

Appendix 6.9.

a. XP-EHH; Ethiopia and Zulu; Affy 6
b. XP-EHH; Ethiopia and Zulu; Omni1
c. XP-EHH; Ethiopia and Zulu; Mega 2.5
d. XP-EHH; Ethiopia and Zulu; H3A
e. XP-EHH; Ethiopia and Zulu; PanAFR
f. XP-EHH; Ethiopia and Zulu; Omni 2.5
g. XP-EHH; Ethiopia and Zulu; Omni5

Appendix 6.10.

a. Tajima's D; Baganda; Affy 6
b. Tajima's D; Baganda; Omni1
c. Tajima's D; Baganda; Mega 2.5
d. Tajima's D; Baganda; H3A
e. Tajima's D; Baganda; PanAFR
f. Tajima's D; Baganda; Omni 2.5
g. Tajima's D; Baganda; Omni5

Appendix 6.11.

a. Tajima's D; Ethiopia; Affy 6
b. Tajima's D; Ethiopia; Omni1
c. Tajima's D; Ethiopia; Mega 2.5
d. Tajima's D; Ethiopia; H3A
e. Tajima's D; Ethiopia; PanAFR
f. Tajima's D; Ethiopia; Omni 2.5
g. Tajima's D; Ethiopia; Omni5

Appendix 6.12.

a. Tajima's D; Zulu; Affy 6
b. Tajima's D; Zulu; Omni1
c. Tajima's D; Zulu; Mega 2.5
d. Tajima's D; Zulu; H3A
e. Tajima's D; Zulu; PanAFR
f. Tajima's D; Zulu; Omni 2.5
g. Tajima's D; Zulu; Omni5

Appendix 7

The number of outlying windows in the intersection between arrays. For a given combination of arrays, indicated by the dark circles at the bottom of the plot, the number of windows in each intersection is depicted by the bar and the number above it.

Appendix 7.1.

a. F_{ST}; Baganda and Ethiopia
b. F_{ST}; Baganda and Zulu
c. F_{ST}; Ethiopia and Zulu

Appendix 7.2.

a. iHS; Baganda
b. iHS; Ethiopia

Appendix 7.3.

a. XP-EHH; Baganda and Ethiopia
b. XP-EHH; Baganda and Zulu
c. XP-EHH; Ethiopia and Zulu

Appendix 7.4.

a. Tajima's D; Baganda
b. Tajima's D; Ethiopia
c. Tajima's D; Zulu

Appendix 8

The distribution of SNP density per window as a percentage of the number of SNPs in the corresponding WGS window, for each accuracy category from the H3A results. The accuracy categories are true positive (TP), true negative (TN), false positive (FP) and false negative (FN).
Plots are grouped into columns according to arrays.

Appendix 8.1.1.

a. F_{ST}; Baganda and Ethiopia; Affy 6
b. F_{ST}; Baganda and Zulu; Affy 6
c. F_{ST}; Ethiopia and Zulu; Affy 6
d. F_{ST}; Baganda and Ethiopia; Omni 1
e. F_{ST}; Baganda and Zulu; Omni 1

f. F_{ST}; Ethiopia and Zulu; Omni 1

Appendix 8.1.2.

a. F_{ST}; Baganda and Ethiopia; Mega 2.5
b. F_{ST}; Baganda and Zulu; Mega 2.5
c. F_{ST}; Ethiopia and Zulu; Mega 2.5
d. F_{ST}; Baganda and Ethiopia; H3A
e. F_{ST}; Baganda and Zulu; H3A
f. F_{ST}; Ethiopia and Zulu; H3A

Appendix 8.1.3.

a. F_{ST}; Baganda and Ethiopia; PanAFR
b. F_{ST}; Baganda and Zulu; PanAFR
c. F_{ST}; Ethiopia and Zulu; PanAFR
d. F_{ST}; Baganda and Ethiopia; Omni 2.5
e. F_{ST}; Baganda and Zulu; Omni 2.5
f. F_{ST}; Ethiopia and Zulu; Omni 2.5

Appendix 8.1.4.

a. F_{ST}; Baganda and Ethiopia; Omni5
b. F_{ST}; Baganda and Zulu; Omni5
c. F_{ST}; Ethiopia and Zulu; Omni 5

Appendix 8.2.1.

a. iHS; Baganda; Affy 6
b. iHS; Ethiopia; Affy 6
c. iHS; Zulu; Affy 6
d. iHS; Baganda; Omni 1
e. iHS; Ethiopia; Omni 1
f. iHS; Zulu; Omni 1

Appendix 8.2.2.

a. iHS; Baganda; Mega 2.5
b. iHS; Ethiopia; Mega 2.5
c. iHS; Zulu; Mega 2.5
d. iHS; Baganda; H3A
e. iHS; Ethiopia; H3A
f. iHS; Zulu; H3A

Appendix 8.2.3.

a. iHS; Baganda; PanAFR
b. iHS; Ethiopia; PanAFR
c. iHS; Zulu; PanAFR
d. iHS; Baganda; Omni 2.5
e. iHS; Ethiopia; Omni 2.5
f. iHS; Zulu; Omni 2.5

Appendix 8.2.4.

a. iHS; Baganda; Omni5
b. iHS; Ethiopia; Omni5
c. iHS; Zulu; Omni 5

Appendix 8.3.1.

a. XP-EHH; Baganda and Ethiopia; Affy 6
b. XP-EHH; Baganda and Zulu; Affy 6
c. XP-EHH; Ethiopia and Zulu; Affy 6
d. XP-EHH; Baganda and Ethiopia; Omni 1
e. XP-EHH; Baganda and Zulu; Omni 1
f. XP-EHH; Ethiopia and Zulu; Omni 1

Appendix 8.3.2.

a. XP-EHH; Baganda and Ethiopia; Mega 2.5
b. XP-EHH; Baganda and Zulu; Mega 2.5

c. XP-EHH; Ethiopia and Zulu; Mega 2.5
d. XP-EHH; Baganda and Ethiopia; H3A
e. XP-EHH; Baganda and Zulu; H3A
f. XP-EHH; Ethiopia and Zulu; H3A

Appendix 8.3.3.

a. XP-EHH; Baganda and Ethiopia; PanAFR
b. XP-EHH; Baganda and Zulu; PanAFR
c. XP-EHH; Ethiopia and Zulu; PanAFR
d. XP-EHH; Baganda and Ethiopia; Omni 2.5
e. XP-EHH; Baganda and Zulu; Omni 2.5
f. XP-EHH; Ethiopia and Zulu; Omni 2.5

Appendix 8.3.4.

a. XP-EHH; Baganda and Ethiopia; Omni5
b. XP-EHH; Baganda and Zulu; Omni5
c. XP-EHH; Ethiopia and Zulu; Omni 5

Appendix 8.4.1.

a. Tajima's D; Baganda; Affy 6
b. Tajima's D; Ethiopia; Affy 6
c. Tajima's D; Zulu; Affy 6
d. Tajima's D; Baganda; Omni 1
e. Tajima's D; Ethiopia; Omni 1
f. Tajima's D; Zulu; Omni 1

Appendix 8.4.2.

a. Tajima's D; Baganda; Mega 2.5
b. Tajima's D; Ethiopia; Mega 2.5
c. Tajima's D; Zulu; Mega 2.5
d. Tajima's D; Baganda; H3A
e. Tajima's D; Ethiopia; H3A

f. Tajima's D; Zulu; H3A

Appendix 8.4.3.

a. Tajima's D; Baganda; PanAFR
b. Tajima's D; Ethiopia; PanAFR
c. Tajima's D; Zulu; PanAFR
d. Tajima's D; Baganda; Omni 2.5
e. Tajima's D; Ethiopia; Omni 2.5
f. Tajima's D; Zulu; Omni 2.5

Appendix 8.4.4.

a. Tajima's D; Baganda; Omni5
b. Tajima's D; Ethiopia; Omni5
c. Tajima's D; Zulu; Omni 5

Appendix 9.

Descriptive statistics for the distribution of SNP density per window in a given array, as a percentage of SNP density in a WGS window. Distributions are grouped into accuracy measure classifications (Acc), from the F_{ST} results. The mean, median and standard deviation (SD) are provided.

Appendix 9.1.

a. F_{ST}; Baganda and Ethiopia
b. F_{ST}; Baganda and Zulu
c. F_{ST}; Ethiopia and Zulu

Appendix 9.2.

a. iHS; Baganda
b. iHS; Ethiopia
c. iHS; Zulu

Appendix 9.3.

a. XP-EHH; Baganda and Ethiopia
b. XP-EHH; Baganda and Zulu
c. XP-EHH; Ethiopia and Zulu

Appendix 9.4.

a. Tajima's D; Baganda
b. Tajima's D; Ethiopia
c. Tajima's D; Zulu

Appendix 10

Values of the Mann-Whitney U statistic (U-stat) with accompanying p values for the difference in the distributions of SNP density percentages between pairs of accuracy measure categories.

Appendix 10.1.

a. F_{ST}; Baganda and Ethiopia
b. F_{ST}; Baganda and Zulu
c. F_{ST}; Ethiopia and Zulu

Appendix 10.2.

a. iHS; Baganda
b. iHS; Ethiopia
c. iHS; Zulu

Appendix 10.3.

a. XP-EHH; Baganda and Ethiopia
b. XP-EHH; Baganda and Zulu
c. XP-EHH; Ethiopia and Zulu

Appendix 10.4.

a. Tajima's D; Baganda
b. Tajima's D; Ethiopia
c. Tajima's D; Zulu

Appendix 11

The number of false negative windows with more than 10 SNPs per window (>10 SNPs) and fewer than 10 SNPs per window (<10 SNPs), the total number of false positive windows before adjacent windows were merged (FN total) and the percentage of false negative windows with a SNP density below 10 (% FN <10 SNPs).

a. F_{ST}
b. iHS
c. XP-EHH
d. Tajima's D

Appendix 12

The concordance between the results of combinations of populations. The heights of the bars and the numbers above them indicate the number of windows identified as selection signals by the given combination of populations or population pairs, specified by the dark circled below the bars.

a. F_{ST}
b. Tajima's D

Appendix 13

Gene and phenotype descriptions for selection candidates from the Ensembl database for each population or population pair.

Appendix 13.1

Baganda

Appendix 13.2

Ethiopia

Appendix 13.3

Zulu

Appendix 13.4

Baganda and Ethiopia

Appendix 13.5

Baganda and Zulu

Appendix 13.6

Ethiopia and Zulu

Appendix 14

Complete lists of 1% outliers for each selection statistic and population or population pair.

Appendix 14.1

F_{ST}; Baganda and Ethiopia

Appendix 14.2

F_{ST}; Baganda and Zulu

Appendix 14.3

F_{ST}; Ethiopia and Zulu

Appendix 14.4

iHS; Baganda

Appendix 14.5

iHS; Ethiopia

Appendix 14.6

iHS; Zulu

Appendix 14.7

XP-EHH; Baganda and Ethiopia

Appendix 14.8

XP-EHH; Baganda and Zulu

Appendix 14.9

XP-EHH; Ethiopia and Zulu

Appendix 14.10

Tajima's D; Baganda

Appendix 14.11

Tajima's D; Ethiopia

Appendix 14.12

Tajima's D; Zulu

Appendix 15

Genes which were not present in a list of previously identified selection candidates, created by modification of the dbPSHP database.

Appendix 15.1

F_{ST}

Appendix 15.2

iHS

Appendix 15.3

XP-EHH

Appendix 15.4

Tajima’s D

Appendix 16

Turnitin report