Exploring the inheritance of complex traits in humans

Peter K Joshi, B.Sc., M.Sc.

PhD – The University of Edinburgh - 2014

Abstract

I explore the genetic and environmental basis of inheritance using modern techniques, in particular high-density genotyping arrays, and older techniques, in particular family history, to explore some longstanding questions about the way we inherit complex traits.

Using pedigree data and the parent-offspring regression technique, I estimate narrow sense heritability (h^2) of human lifespan in 20th Century Scotland as 0.16, lower than commonly cited studies in other populations. I also observe similar concordance between spouses as between parents and offspring - suggesting my estimate of heritability may include significant within-family environment effects and thus should be considered an upper bound.

Using genome-wide array data to identify runs of homozygosity, from 150 cohorts across the world and up to 350,000 subjects per trait, I show that cognitive function and body size are associated with the total length of genome-wide runs of homozygosity. Contrary to earlier reports in substantially smaller samples, no evidence was seen of an influence of homozygosity on blood pressure and low-density lipoprotein (LDL) cholesterol, or ten other cardio-metabolic traits. An association between genome-wide homozygosity and complex traits arises due to directional dominance. Since directional dominance is predicted for traits under directional evolutionary selection, this study provides evidence that increased stature and cognitive function have been positively selected in human evolution, whereas many important risk factors for late-onset complex diseases have not.

The analysis of less common single nucleotide polymorphism (SNP) variants in genome-wide association studies promises to elucidate complex trait genetics but is hampered by low power to reliably detect association, whilst avoiding false positives. I show that addition of 100 population-specific exome sequences to 1,000 genomes global reference data allows more accurate imputation, particularly of less common SNPs (minor allele frequency 1–10%). The imputation improvement corresponds to an increase in effective sample size of 28–38%, for SNPs with a minor allele frequency in the range 1–3%.

Inheritance of complex traits remains a field wide open for discovery, both in determining the balance between nature and nurture and discovery of the specific mechanisms by which DNA causes variation in these traits, with the prospect of such discoveries illuminating biological pathways involved and, as knowledge deepens, facilitating prediction.

Declaration

(a) this thesis has been composed by me, and

(b) the work is my own except as indicated below

(c) that the work has not been submitted for any other degree or professional qualification.

……………………………………….

Peter Joshi 14 January 2015

Assistance

Drafting

My supervisors commented on drafts of all chapters in this thesis and suggested rewordings, additional references and modifications to the tables and figures.

ORCADES and CROATIA-Korcula data

The phenotype, pedigree, called array genotyping and sequence data for these studies were provided to me by Jim Wilson, for subsequent quality control and analysis by me.

Chapter 2 Heritability of Human Longevity

The EASTER project was conceived, designed and carried out by me, except for the assistance I received in the collection of data. Michael Tobias of Hymans Robertson wrote the extraction scripts, which created the original candidate list of rare names, these were then looked up by Jordan Irvine, Gaetano Donato and Holly Trochet. Holly Trochet designed, implemented and wrote the description of the occupational and social level classifications shown in Supplement 2.5.1, which were not taken forward into main results, and parts of the description data gathering method (2.1 & 2.2.1).

Chapter 3 Optimisation of PLINK Runs of Homozygosity calling

The need for to determine a suitable approach for multiple array studies and the use of HapMap data to do so were suggested by Jim Wilson.

Chapter 4 The effect of homozygosity on 16 complex trait

This work was a collaboration – between the four principal authors (Jim Wilson, Tõnu Esko, Ozren Polašek and myself) and the participating cohorts. Based on previous published work, I designed and wrote the analysis plan and cookbook for its implementation, with testing carried out by Toñu and several cohort analysts. Meta-analysis was carried out by me and I prepared the first draft of the manuscript, which was then subsequently redrafted based on the comments of all authors, especially Jim Wilson.

Chapter 5 Using local exome sequences to impute hidden variants and increase power of Genome Wide Association Studies

This work was suggested by Pau Navarro and Chris Haley. James Prendergast and Ross Fraser prepared the exome sequence genotypes and provided me with the variant call format file. James wrote the exome sequence calling section of the methods. Pau guided my approach and commented on my experiments, as I designed them, and all authors (and in due course peer reviewers) commented on the manuscript prepared by me.

Contents

Figures and Tables

Acknowledgements

I would like to thank

Michael Tobias, Holly Trochet, Gaetano Donato and Jordan Irvine for gathering the family history data for the Easter study and Douglas Anderson of Hymans Robertson for providing the funding for the data gathering;

The International HapMap consortium and 1,000 Genomes projects for making their data publicly available;

All those who have contributed to the creation of the population genetic studies ORCADES and CROATIA-Korcula, especially the study subjects, the study funders and Jim Wilson, Harry Campbell, Caroline Hayward, Alan Wright, Kay Lindsay, James Prendergast, Ross Fraser and Ruth McQuillan;

Ross Fraser for setting up and maintain the computing resources, which were vital to my analyses;

All the participating cohorts in the ROHgen consortium;

The many authors of the computing tools used by me: Shaun Purcell for PLINK, Linus Torvalds for the whole Linux project, the R project team for the R computing language and the packages used by me, especially Reshape by Hadley Wickham, Rmeta by Thomas Lumley and GenABEL by Yurii Aulchenko;

The Medical Research Council for funding my scholarship; and

My supervisors Pau Navarro, Chris Haley and Jim Wilson, for granting me access to their data, suggesting the experiments, guiding me through their conduct, helping me interpret the results and providing comments on this thesis.

I would also like to thank all my supervisors, especially Jim Wilson, for their constant encouragement, support and good humour over the last three years.

Glossary

Allele	One of a number of genetic variants present in a population
Causal Variant	A genetic variant that has a causal effect on the complex trait (as opposed to one that is merely associated with variation in the trait, due to LD)
Complex Trait	An observed characteristic that is affected by a large number of genetic factors and the environment
Dominance	The process by which a heterozygote is not precisely intermediate to the two homozygous forms, but is more similar to one homozygous form than the other. (eg brown eye colour is dominant: a heterozygote for blue/brown eye alleles has brown eyes, rather than an intermediate colour)
Directional Dominance	A consistent genome wide pattern of dominance favouring either an increase or decrease in a complex trait
Epistasis	The process by which the effect of one gene is affected by variation in another gene, i.e. interaction between genes.
F	The measure of the degree of inbreeding in an individual $0 \leq F \leq 1$
F_{ROH}	The genomic measure of F based on

	observed runs of homozygosity.
Single nucleotide polymorphism (SNP)	A single base of the DNA genetic code that varies between individuals.
Genome	The complete (~3 billion base pair) genetic code of an individual
Genome Wide Association Study (GWAS)	A study of the association between individual genetic variants (typically SNPs) across the genome and a trait
Genomic Relationship Matrix (GRM)	A measure of the relatedness between individuals based on genotype
Genotype	The combination of genetic variants (alleles) specific to an individual. In human autosomal chromosomes, individuals have two alleles at each locus.
Genotyping array	An assay platform that measures genotype at many (often ~100k – 3M) variables sites (often SNPs)
Heterozygote	An individual whose genotype is two different alleles at a genetic locus.
Homozygote	An individual whose genotype is two of the same alleles at a genetic locus.
Identity-by-descent (IBD)	A homozygous genotype arising due to identical alleles being inherited from a single common ancestor
Imputation	Estimation of genotypes using statistical inference techniques and knowledge of linkage dis-equilibrium between

	genotypes at known and unknown loci
Heritability	The degree to which a trait is determined genetically, as measured by the proportion of variance attributable to genetic causes. Heritability is population specific as both genetic and non-genetic variation in a trait is specific to a population.
Linkage Disequilibrium (LD)	The association in genotype at different loci, often due to their genomic proximity and co-inheritance.
Minor Allele Frequency (MAF)	The frequency of the less common allele in a population
Run of Homozygosity (ROH)	A contiguous region of the genome in which all observed genotypes are homozygous, the region, including intermediate unobserved loci, is inferred to be IBD.
Summed Runs of Homozygosity (SROH)	The sum of the length (in base pairs) of all observed ROH in an individual.

Chapter 1 Introduction

1.1 Complex Traits

Complex traits are observed characteristics which are affected by a large number of inherited genetic factors and the environment, as well as interactions between them[1]. Human height is a classic example of such a trait [2], but susceptibility to common late-onset diseases, which are the main causes of death in the West, and many disease risk factors such as blood pressure and cholesterol levels are also complex traits [3]. The propensity of such traits, especially height, to run in families may have been obvious for centuries and was first quantified by Galton as long ago as 1886 [4], but it was not until 1918 that Fisher first described how the observed distribution of such traits could arise despite particulate Mendelian inheritance from the small contribution to the phenotypic variance of many (unknown) inherited genetic variants[5].

Thus, there has long been an understanding of some aspects of the inherited (or genetic) basis of such traits. In particular, the extent to which traits run in families has been formalised mathematically, into a measure called (narrow-sense) heritability (h^2). h^2 measures the proportion of phenotypic variance explained by additive genetic factors and a variety of techniques have been developed to estimate heritability from pedigree data, for example twin concordance, or the amount of variance explained within as opposed to between families [6] .

Equation 1-1

$$h^2 = \frac{V_A}{V_P}$$

Where V_A = additive genetic variance and V_p is total phenotypic variance[6].

However, although the amount of genetic variation could be estimated and such estimates were replicated, the underlying genetic variants affecting the traits were unknown.

The first breakthrough in identifying clinically important genetic variants was linkage analysis, which looked at how a trait and a genetic marker co-segregated, or were inherited together, within families [7]. However, although more than 12,000 monogenic disease loci have now been identified [8] , until the 1990s the success in identifying the genetic basis of diseases had not extended to complex disease but instead had mainly been restricted to rare, early onset diseases and to single genes with very large, often devastating, effects [3].

However, with the advent of the human genome project and its completion in 2003, mapping the ~3 billion base pairs of the human genetic code and making them widely available to researchers, geneticists promised society a revolution in the diagnosis and treatment of complex disease and, even more excitingly, in its prediction and prevention[9]. The subsequent development of genotyping arrays, whereby 100k-1m single-nucleotide polymorphism (SNP) markers could be genotyped simultaneously across a patient's whole genome, facilitated substantial effort in Genome-wide Association Studies (GWAS). In a GWAS, a phenotype (such as disease status or a quantitative trait) is tested across a study population for an association with the genotype of each of many hundreds of thousands of SNPs across the genome. GWAS researchers hope that a causal allele, even if not on the SNP panel adopted, is in linkage disequilibrium (LD) with a SNP measured, thereby causing an association between some genotyped SNPs and the trait under study and thus locating a causal region of the genome[3]. Whilst the work scored many successes, [10], it tended to find that susceptibility to common diseases was caused by many genetic variants of small effect. Even when aggregated, the causal alleles discovered only explained a small proportion of the total known genetic variance [3].

This in turn led to larger and larger studies with greater statistical power, discovering variants of even smaller effect size [10] [11].

1.2 Missing Heritability

Despite ever larger studies identifying more and more variants of small effect, most of the expected heritability of complex traits based on pedigree studies has yet to be pinned down to specific variants. For example, whilst 180 SNPs have been found that affect human height in a study of 180,000 people, those SNPs only explain one fifth of the accepted heritability [2], exemplifying the problem known as missing heritability [12]. As a result the allelic architecture underpinning most complex traits remains a matter of controversy[13].

I suggest that **missing** narrow-sense (additive) heritability must arise in one of two ways (perhaps in combination): either the accepted consensus estimates of narrow-sense heritability are over-estimates or there are genetic variants whose additive contribution has not yet been robustly identified. This could be due to causal regions not yet being identified at all, or incomplete linkage of the true causal variant with an identified marker in the region. I exclude systematic under-estimation of effect sizes, due to the known tendency in the opposite direction: the winner's curse [14]. Overestimation of additive heritability in humans might arise due to dominance, epistasis, gene-environment interactions or common environmental confounding [15]. However, the general predictive success of the breeder's equation, which uses heritability to estimate the effect of artificial selection[16], suggests that, at least in some animals, additive genetic variance is similar in scale to that estimated from pedigrees. At the same time, we must recognise that results from domesticated animals, where environmental confounding can be controlled by experimental design, cannot be directly interpreted in natural human populations. On the other hand, if present heritability estimates are broadly correct, there are many unidentified additive genetic variants contributing too little genetic variance for existing GWAS power to detect them. Such variants must either be only weakly linked to the markers being analysed, uncommon (in itself likely to cause weak LD with common markers), or have small effects, as GWAS should have already found common variants of large effect[12]. Genetic research can thus productively attack this problem at both ends: improving top-down estimates of heritability and identifying more and more causal variants.

Fairly recently, researchers showed that around half of the heritability of human height should be explicable by common markers of the type present on modern genotyping arrays, if only studies were large enough to attain sufficient power to reliably detect the small effect sizes concerned [17]. The study estimated heritability based on a seeming oxymoron – the genomic relationship matrix (GRM) amongst unrelated individuals. The contradiction is resolved by understanding that apparently unrelated individuals are distantly related, and, importantly, the degree of relationship varies. Furthermore amongst such subjects, relatedness at one part of the genome should not be informative of relatedness at another part: restricted maximum likelihood (REML) estimates of heritability thus only reflect variance explained by relatedness locally in the genome, rather than allelic correlations at distant loci inferable from local relatedness, as is the case in higher kinship studies. This research thus suggested that increasing statistical power from increased sample size using existing GWAS arrays, and otherwise conventional techniques, would yield more and more common variants, but explained heritability would still only reach at most half of estimated total heritability.

The above relationships can be conveniently summarised in the following equation (following Zaitlen *et al.*'s nomenclature) [18]

Equation 1-2 The postulated transitive nature of narrow sense heritability estimates

$$h^2_{ped} > h^2 > h^2{}_g > h^2{}_{gwas}$$

Heritability is the proportion of variation in a trait for a population that is explained by genetic differences amongst the population. This is known as broad sense heritability and denoted H^2. A specific portion of that heritability is additive – i.e. is transmitted between generations, and is therefore of particular interest. This is called narrow-sense heritability and denoted $h^{2.}$

h^2_{ped} : narrow sense heritability estimated using familial resemblance techniques

h^2 : true (unknown) narrow sense (additive) heritability

$h^2{}_g$: heritability measured using a genomic relationship matrix, using common SNP genotyping arrays, in an unrelated population

$h^2{}_{gwas}$: heritability measured using variance explained by well-established GWAS loci

I.e. we postulate that family studies (h^2_{ped}) over-estimate narrow-sense (i.e. additive) heritability. Such studies typically look at how relatives resemble each other more than the general population and partition variance in the trait into within family and between family components, in lay terms, the studies look at how strongly a trait (eg height) runs in families (i.e. families as a whole being short or tall). Using these techniques can lead to over-estimation of narrow sense heritability for two broad reasons. Firstly, total (broad sense) heritability, includes genetic variance which arises from non-additive causes, such as dominance variation and epistasis. Dominance variation, for example, gives rise to sibling resemblance. It can be difficult to distinguish total heritability from additive heritability and this can lead to over-estimates of narrow-sense heritability. The second broad reason that family resemblance studies may over estimate heritability (in both the broad and narrow sense) is that families typically share a common environment (eg socio-economic status) and this can lead to familial resemblance (eg educational attainment) that is not a result of genetics.

GRM studies of unrelated subjects under-estimate heritability (h^2_g) as they (intentionally) fail to capture loci not in LD with the array SNPs. And finally, variance explained by discovered GWAS SNPs is lower still, due to insufficient power to identify truly associated array SNPs , beyond reasonable doubt.

Precisely where true h^2 lies within the interval between h^2_{ped} and h^2_g remains a matter of speculation, with Zuk *et al.* postulating a role for epistasis [11] inflating h^2_{ped}, whilst Yang *et al.* suggest h^2_g 's exclusion of rare variants may be the principal explanation for the gap between h^2_{ped} and h^2_g. In any case, for many complex traits around half of estimated h^2_{ped} remains unexplained by h^2_g [17] and thus much more work is needed. To refine these limits we should seek to increase h^2_g through use of denser arrays or sequence data, with more rare variants, and increase h^2_{gwas} through larger studies and improved imputation. At the same time, improvements in estimates of h^2_{ped} are needed, by narrowing confidence intervals and avoiding biases such as environmental confounding, and interactions such as dominance and epistasis.

1.3 Better estimating (narrow-sense) heritability using h^2_{ped}

Although the mathematical definition of h^2 is precise (Equation 1-1) and its consequences – the part of genetic value passable between generations [6] – clear, measurement of heritability in human populations remains controversial. One source of controversy is practical: large studies are needed to estimate variance components with any accuracy. The other source of controversy is the methodology itself. Studies of inheritance are known to be subject to confounding of genetic and environmental effects, especially due to common familial environment, whilst twin studies assume that shared environment is equal for monozygotic and dizygotic twins [18]. Well powered analysis of inheritance, particularly if it avoided the twin design and allowed probing of non-genetic familial resemblance, is thus still worthy of interest, especially if the trait is an interesting one.

A large study of the inheritance of human lifespan between parents and offspring fits the bill and is a study design which avoids (non-additive) dominance variation unintentionally giving rise to inflation of additive estimates of variance components [6]. Researchers commonly cite an estimate of 0.25 for the (h^2) heritability of human longevity, despite the study concerned not being well-replicated, not quoting confidence intervals and using twins [19]. A useful study could consider intra-marriage correlations as well as intergenerational ones, with the potential to confirm or challenge the genetic basis of missing heritability, albeit with other experimental design limitations, such as the effect of assortative mating. Such a study would give useful new estimates of additive genetic variation in one of the traits of greatest interest to people – the length of their own lives.

1.4 Non-additive genetic variation

Much of the focus of genetic research, especially in animals, is on additive variance, partly as this is the variation upon which selection can act [6]. Nonetheless many genetic mechanisms exhibit dominance or recessivity [8] and epistasis, as well as gene-environment interaction, which will all contribute to inheritance of complex traits.

Although apparently non-additive variation, such as recessive effects, do contribute to additive variation [6] and can thus be identified using an additive model, it is interesting to try to discern such effects more directly. However simply extending GWAS techniques, where hundreds of thousands of genetic markers are tested to other models would multiply the problems of multiple testing. Directional dominance, where there is a consistent genome-wide pattern of dominance is of particular interest, as its existence has long been speculated upon based on evolutionary considerations and evidence from some of the earliest genetic studies using pedigrees and measures of inbreeding [20]. Modern molecular techniques using genomic runs of homozygosity (ROH) [21] allow the study of directional dominance even in outbred populations, where consanguinity is not practiced, akin to the use of the GRM to estimate heritability in unrelated populations. Determining the existence and extent of directional dominance thus provides an interesting insight into the evolution and inheritance of complex traits.

1.5 Imputation

Whilst understanding of the inheritance of complex traits can be furthered by top down inferences (such as the estimate of h^2) or genome-wide analyses such as directional dominance, such analyses provide little or no information on the biological mechanisms underpinning the trait. Bottom up understanding of genetic inheritance stems from the identification of genetic variants associated with phenotype. Whilst, linkage between (phenotypic visually observable) markers and other traits was observed as early as 1905 [22] , it took many more years before individual SNPs could be genotyped, and the subsequent developments of Sanger and next generation sequencing before the reference human genome could be assembled, facilitating the development of genotyping arrays and whole genome sequencing, finally enabling large scale GWAS – genome-wide scans across general populations for the association between SNPs and complex traits.

Although whole genome sequencing continues to fall in cost, most GWAS presently have array data as their base genotypes. Whilst modern arrays typically assay 100k-1m SNPs, they only capture a small proportion of the more than 40 million SNPs found at a global population level [23]. A genetic study using array data alone therefore loses power due to incomplete LD between the array SNP and the causal variant. At the same time, different studies have used different genotyping arrays, which have measured different SNPs. Combining studies is only straightforward for SNPs that have been genotyped on all arrays – therefore substantially reducing the number of SNPs that can be analysed, or reducing the number of studies that can be combined. Using genotype array data alone thus has two drawbacks – reduced power and difficulties in meta-analysis.

Imputation, a process by which genotyping array data is used to estimate genotypes that are not genotyped on the array, can overcome these drawbacks. Firstly it can increase power, by estimating the latent genotypes accurately. Secondly the imputed genotypes can be used as a common panel across all studies wishing to participate in a meta-analysis [24], an approach, now regular practice [2] enabled by imputation. Thus, improving the accuracy of imputations offers the prospect of improving the ability of meta-analyses to identify causal DNA variants affecting the inheritance of complex traits and illuminating the biological processes involved.

1.6 Conclusion

As suggested above, I saw the opportunity to research the inheritance of complex traits in three seemingly diverse areas

- using family history to measure the extent to which human longevity runs in families and investigate whether this is nature or nurture;
- using genome-wide array data to determine the extent of distant parental relatedness and its effect on complex traits; and
- improving the quality of estimation of genotypes at loci that have not been directly genotyped, to facilitate the discovery of causal variants underpinning complex traits.

My argument is simple: much remains to be discovered about the inheritance of complex traits in human. A combination of long established methodology, unprecedented amounts of genomic information now available and new techniques, offers the prospect of insights into both bottom up and top down understanding of how inherited DNA contributes to who we are.

Chapter 2 The inheritance of human lifespan in 20th Century Scotland

Abstract

We estimated the narrow sense heritability (h^2) of human longevity, conditioned on living beyond the age of 42. Data on age at death were drawn from Scottish public records for 2,984 individuals with rare surnames born in eastern Scotland around 1900, and their parents: the EASTER study. We also collated 1,854 Orcadian parent-offspring trios, where the offspring was born between 1880 and 1920.These trios were ancestors of subjects in the Orkney Complex Disease Study, ORCADES.

Using the correlation between parent and offspring, we estimate h^2 for longevity as 0.166 (95% CI 0.126-0.206) in these populations, which is somewhat lower than previous recent estimates (0.22-0.35) in other populations, which we suggest may be inflated due to methodological problems in the other studies, or the special nature of the other populations under study. We also note that our measured correlation between spouses is very similar to that between parents and offspring. We infer that latent shared within family environment effects may be included in our estimate. We therefore believe our estimates should be considered as upper bounds for heritability of lifespan in our population.

This is the first large non-twin based study of the inheritance of human longevity in a general population, living in conventional developed 20th century western society. Our results suggest the genetic heritability of longevity may be somewhat lower in these societies than suggested by previous studies elsewhere.

2.1 Introduction

Human longevity is (literally and arguably metaphorically) the ultimate human phenotype and is a consequence of many intermediate complex trait and disease phenotypes and other genetic and environmental factors, including chance. Research into the inheritance of longevity has a long history. The first such paper appears to be from 1899 by Beeton and Pearson, who analysed the correlation between the lifespans of aristocratic English fathers and sons [25] and a wide range of studies were reviewed in 1964[26]. Often these studies relate to specific and unusual populations and the people under study lived a long time ago. In more recent years, whilst there have been a large number of successful genome-wide association studies [27] and heritability estimates [28] for disease incidence that are the common causes of death in the developed world, the heritability of longevity has been less conclusively analyzed, for a number of reasons.

Whilst age at death is straightforward to measure objectively, the preceding process obviously takes a lifetime. Thus for studies using healthy subjects, the delay to the subjects' deaths will normally exceed the time horizon of the researcher. One obvious approach is thus to look at subjects already dead. Although genotypic information is unlikely to be available, public or other records can be used to identify family members (especially twin pairs) and determine their age at death, and thus investigate the inheritance of lifespan, using established techniques [6], although secular changes in longevity can complicate analysis.

In recent years, only a few sufficiently large studies have estimated the heritability of human longevity (measured by age at death or susceptibility to death), with estimates in the range of 0.22-0.5 [29] [19] [30] [31] [32]. The studies can be conveniently grouped into two categories – twin studies and studies of other kinds of relative pairs. Twin-based studies typically compare the correlation in a trait of monozygotic and dizygotic twins, and attribute the excess correlation in monozygotic twins as being solely due to the complete genetic sharing. Such estimates will include dominance and epistatic effects and thus estimate broad sense heritability. They also implicitly assume that environmental similarities between monozygotic and dizygotic twins are the same. The largest twin study of human longevity, using age at death as the phenotype, has estimated heritability of 0.25, but with unclear standard errors [19].

A second approach is to look for phenotypic correlation amongst siblings or parents and offspring. These studies suffer the disadvantage of not necessarily being able to control for within family environment, but for parent offspring designs, the approach avoids dominance and common maternal and childhood environment effects, whilst introducing inter-generational effects and within family effects. Heritability of longevity (+/- 1 SE) of 0.25 (+/- 0.05) was estimated amongst the Old Order Amish, an isolated population with an unusual and communal living environment, that had been stable over generations [30]. Two other recent studies have estimated the heritability of longevity in the range 0.35-0.5, using more difficult to interpret statistics of relative mortality risk, using complex methods [32] [31].

Rather than estimating heritability of lifespan, other researchers have considered long-livedness, perhaps beyond age 90 or 100, as a binary trait. These studies have sometimes [33,34], but not always [35] shown excess concordance between blood relatives in long-livedness, relative to controls. However, we shall not consider such approaches further here, as they lend themselves less well to the estimation of heritability of lifespan *per se*.

It is therefore of interest to estimate accurately the narrow sense heritability of longevity, defined simply as age-at-death, in social settings more typical of Western European life in the 20th Century, using parent–offspring correlations.

The ScotlandsPeople Centre (National Records of Scotland and the Court of the Lord Lyon 2012) in Edinburgh has publicly available computer indices of birth, death, and marriage records from the year 1855 to the present. Records can be searched by name and year and visually inspected on the computers at the Centre itself. Importantly women's maiden names are indexed, which assists greatly in the identification of correct parents. Scotland is therefore an excellent place to research the genetics of longevity in modern populations, using pedigree methods.

The Orkney Complex Disease Study (ORCADES) is a family-based, longitudinal community study of the genetics of complex traits, based in the Orkney Isles in Scotland [36], for which we have gathered pedigree information directly from participants, supplemented by data from the ScotlandsPeople Centre.

We have conducted two population-based studies, looking at the inheritance of longevity by measuring the correlation in lifespan between offspring born around 1900 and that of their parent. The first study – EASTER- was created entirely for this project and looked at people with rare surnames born between 1892 and 1910, who died in Fife or Angus (two contiguous counties on the eastern shores of Scotland). We specifically selected offspring with rare surnames to facilitate tracing their parents' death records. The second study looked at ancestors of ORCADES subjects born between 1880 and 1920.

2.2 Method

2.2.1 EASTER Data Gathering

We ran an automated web extraction on the Scottish Records Office's online indices to generate lists of names of people who had died in the regions of interest after the age of 42. The age of 42 was chosen to minimize the number of deaths from either World War and from accidents. These were filtered for people with rare surnames to facilitate the tracing of parents. "Rare surnames" were names with fewer than 150 matching electronic earlier male records with the same surname - i.e. possible father death records for the death in question. The death records for these candidate subjects were then looked up manually and parental details captured. The online database was then searched for other deaths with the same surname and the resultant candidate parent death records were reviewed manually to find the matching parent. To ensure that we were matching the correct parents to the correct children, we cross-checked the parent forenames, spouse names and occupations listed on both the child and parent death certificate, allowing some leeway for spelling of surnames, as that sometimes changes from record to record for the same individual. Matching on this basis, although manual, was almost always clear and unambiguous, with any doubt usually being resolvable by further cross-reference to the marriage certificate. Records were skipped if tracing proved too difficult, if the individual's cause of death was suicide, war death, or accident, or if the individual was identified as a sibling of someone already in the dataset, or the analyst did not believe there was an unambiguous match. For the subject, we gathered dates of birth and death, whereas for the parents we gathered age-at-death. We also gathered socio-economic information in the form of the subject's and father's occupations and the usual residence at death.

For the EASTER study, they were 4,385 candidate trios. However, after, exclusions, for example due to the inconvenience of not being truly a rare name, or due to sibship, or the offspring or either parent having died by accident, war or suicide, and excluding any trios that were incomplete, usually due to one parent's death record not being found, there were 3,266 trios, with offspring born 1892-1910 available for analysis.

2.2.2 ORCADES Data Gathering

The initial data available for the ORCADES study was in a different format: pedigree information is available for 2,124 ORCADES subjects, most of whom are still alive. However, age at death is recorded for subjects' ancestors. We considered potential trio offspring up to and including as the subjects' grandparents (and thus potential parents in the trios from the subjects' great grandparental generation). 8,470 unique trios were identified. However, in many cases at least one member of the trio was still alive, after excluding trios where members were alive, there were 5,650 trios. To increase comparability with the EASTER data and reduce the effect of any secular trends in longevity, the ORCADES study was then restricted to trios, whose offspring were born 1880-1920, leaving 1,858 for analysis. As cause of death was not recorded, we did not exclude individuals who died through suicide, accident or war.

Of the 3,266/1,858 EASTER/ORCADES trios, a further 282/200 were excluded because one trio member had died before the age of 42. The final number of trios was 2,984 (1,630: 1,354 with Male: Female offspring)/1,658 (832: 836 with Male: Female offspring) for the EASTER/ORCADES cohorts respectively. Offspring were typically born in 1900 and on average died in the late 1970s. Parents were typically born in the 1860s and died on average around 1940 (Table 1 and Table 2). Figure 1 plots the distributions of age at death in these datasets.

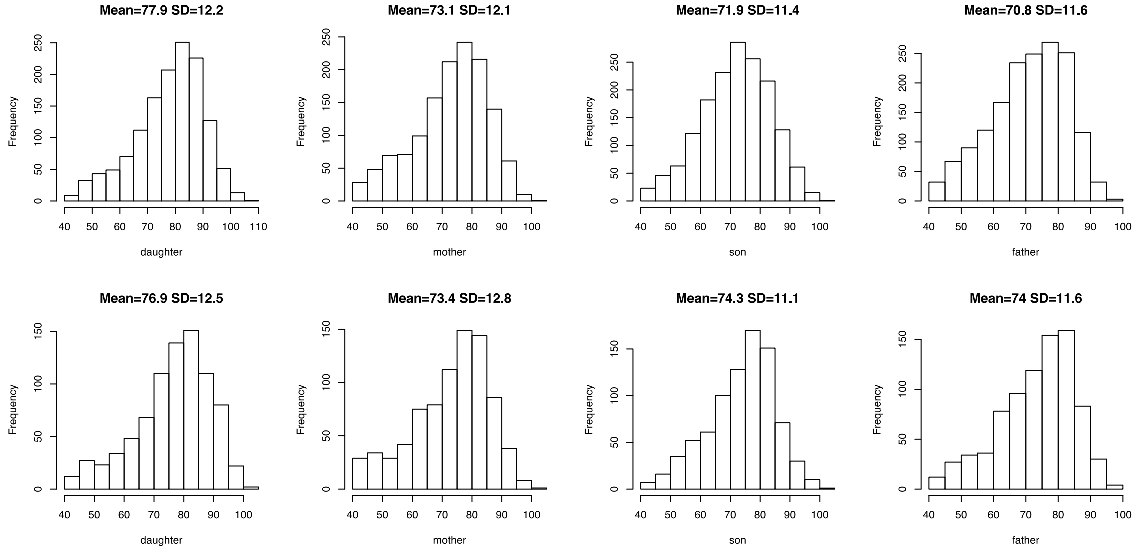
Table 1 Mean, Median and Standard Deviation of Year of birth, Year of death and Age at death for the EASTER study

		Sons	Daughters	Fathers	Mothers
	Count	1630	1354	2984	2984
Age at death	Mean	71.9	77.9	71.1	72.5
	Median	72.8	79.8	72.5	74.5
	SD	11.4	12.2	11.6	12.2
Year of birth	Mean	1902	1902	1868	1870
	Median	1903	1902	1868	1871
	SD	5.31	5.25	9.08	8.44
Year of death	Mean	1974	1979	1939	1943
	Median	1975	1981	1940	1944
	SD	12.8	13.6	15.3	14.9

Table 2 Mean, Median and Standard Deviation of Year of birth, Year of death and Age at death for the ORCADES study

		Sons	Daughters	Fathers	Mothers
	Count	832	826	1658	1658
Age at death	Mean	74.3	76.9	74.2	74.1
	Median	76	79	76	77
	SD	11.1	12.5	11.4	12.4
Year of birth	Mean	1901	1900	1864	1868
	Median	1901	1901	1864	1868
	SD	11.3	11.5	13.6	13.1
Year of death	Mean	1975	1977	1939	1942
	Median	1975	1978	1939	1942
	SD	15.6	17.8	16.9	17.8

__Figure 1__ Distribution of Ages at death for subjects included in the final study dataset



Row 1 is the EASTER study
Row 2 is the ORCADES study

All statistical analyses were carried out using the software package R [37] and the library RMETA [38]. For each cohort, the correlation in lifespan was calculated between offspring of each sex and each parental sex. A combined correlation was calculated for both sexes and both parents, from the correlation in the residuals, for each generation separately after fitting sex as a fixed effect (i.e. residuals from a linear model: age-at-death ~ sex).

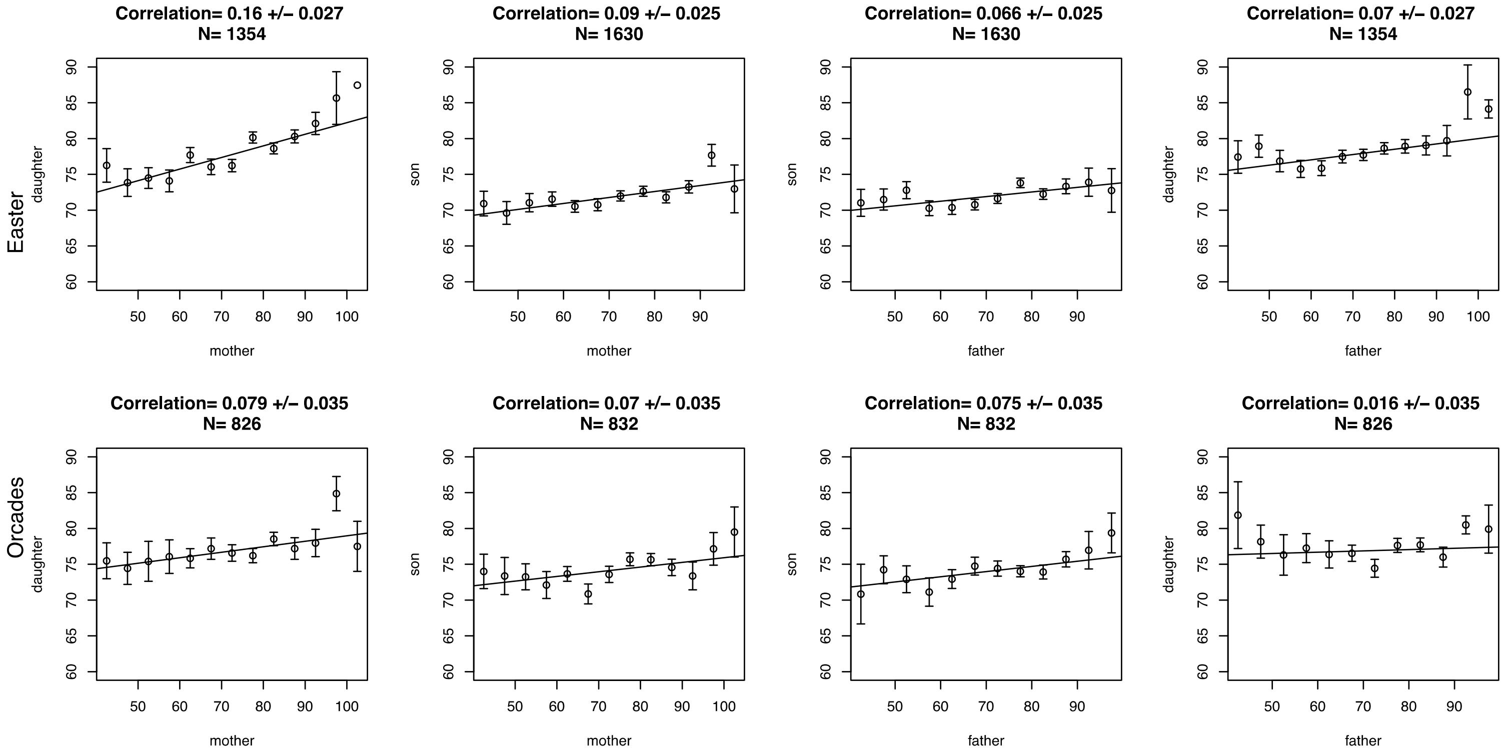
Initially consideration was given to fitting further covariates available in the EASTER study, such as occupational class and location. However, the variance explained by these covariates was small and including them in the analytical model made little difference to the results (supplement 2.5.1). As a consequence, the decision was made not to fit available socio-economic covariates (in EASTER), to give a more consistent protocol across the two studies.

The estimated correlations from the two studies were combined using the Inverse variance method and Fisher's z-transformation of correlations. Narrow-sense heritability (h^2) was estimated as twice the correlation between offspring and parent.

2.3 Results

We estimated narrow sense heritability using parent-offspring regression. Figure 2 shows the relationship for each sex separately. Visual inspection suggests the relationship is broadly linear, although some graphs appear to show that a parent dying past age ninety was associated with greater offspring longevity than the linear regression implied.

Figure 2 Regression of parent age at death on offspring age at death, split by sex



To aid visibility (and in the graph only) parental ages have been split into 5-year age brackets.
Offspring age at death shown is the mean in parental age bracket.
Error bars are 1 standard error of the offspring mean in each bracket.

The correlations between parent and offspring lifespan (95% CI) ranged from 0.016 (0-0.0859) (ORCADES daughter-father) to 0.1604 (0.1068-0.214) (EASTER daughter/mother), as shown in Table 3. The correlation between the fathers' and mothers' lifespan was of a similar magnitude to that with their offspring. Father-mother correlation in lifespan was 0.075 (n=2,983,95% CI 0.039-0.112) for EASTER and 0.051 (n=1,656, 95% CI 0.002-0.100) for ORCADES.

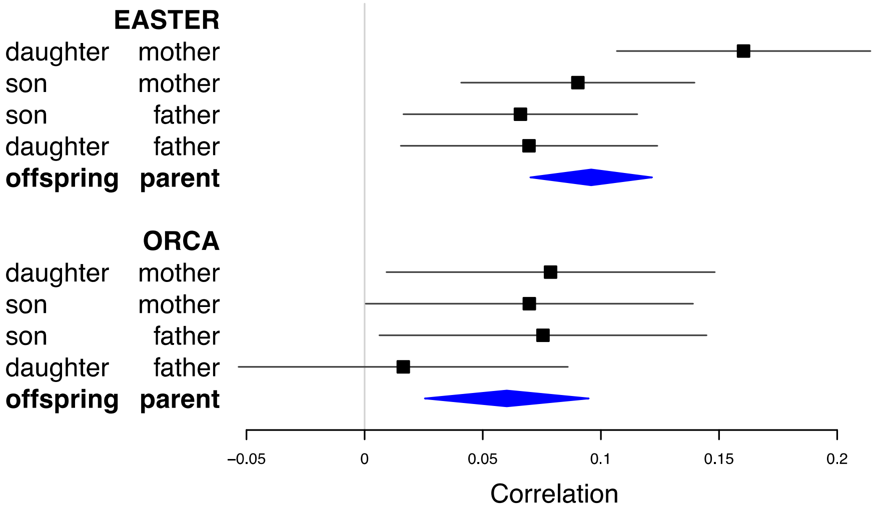
Table 3 Correlation in Parent and Offspring ages at death

Study	**Offspring**	**Parent**	**Number of Duos**	**Correlation in ages at death (1)**	**95% CI**
	daughter	mother	1354	0.1604	0.1068-0.214
	son	mother	1630	0.0902	0.0408-0.1396
EASTER	son	father	1630	0.0659	0.0165-0.1153
	daughter	father	1354	0.0696	0.0154-0.1238
	offspring	**parent**	**5968**	**0.0959**	0.0701-0.1217
	daughter	mother	826	0.0787	0.0093-0.1481
	son	mother	832	0.0697	0.0005-0.1389
ORCADES	son	father	832	0.0755	0.0063-0.1447
	daughter	father	826	0.0163	0-0.0859
	offspring	**parent**	**3316**	**0.0601**	0.0255-0.0947

For the two rows where all offspring are regressed on both parent, the correlation is in age at death residuals having fitted sex as a fixed effect (i.e. age-at-death ~ sex) to each generation.

Figure three illustrates these results graphically. The observed correlations across the different parent and offspring sexes and across studies appear broadly mutually consistent, although the correlation between EASTER mothers and daughters 0.16 is rather higher than the overall correlation 0.08. Nonetheless, in meta-analysis testing for heterogeneity across the different sex results and studies, the observed correlations appear to be mutually consistent, p-value = 0.071.

Figure 3 Forest plot of observed correlation of parent and offspring ages at death and 95% CI



The combined estimate across the two studies of the correlation between parent and offspring age at death is 0.0831 (95% CI 0.0629-0.103). The combined estimate of heritability is thus 0.166 (95% CI 0.126-0.206). There was no statistically significant (p>0.05, even before allowing for multiple testing) evidence of differences in correlations between parents and offspring, for ORCA & EASTER, father & mothers, intra-sex & inter-sex and sons & daughters (p= 0.100,0.066,0.146,0.516 respectively).

Across all the offspring groupings, we do not believe that our study design is significantly influenced by unintentional under or over sampling parts of the distribution of offspring age at death. Firstly we are interested in the regression, not the mean. A shift in the mean of the explanatory variable should not affect the measured slope. Whilst truncation might affect correlations, we note the stable standard deviations across our generations, sexes and studies. Finally, the empirical evidence does not suggest an age-related slope – both from visual examination of the betas in Figure 2 and the results in Supplementary note 2, where removal of the more extreme ages, often increasing the mean age at death, had little effect on observed correlation.

2.4 Discussion

Our study is the first to study of sufficiently large scale to establish reasonably tight estimates of the heritability of lifespan for populations with normal living environments in 20th Century Western Europe. Using regression on over 8,300 parent-offspring relationships, for children born in Scotland around 1900, we estimate the (narrow-sense) heritability of human lifespan in that population as 0.166 (95% CI 0.126-0.206). The parent-offspring method estimate has the advantage of avoiding common maternal, and dominance effects, whilst within family environmental effects may still be (wrongly) captured in the estimate [18]. Our estimate is lower than the most commonly cited, 0.25, in a Danish twin study [19] and other recent estimates [30] [31] [32]. Whilst accepting no single study should claim to be definitive, we believe our study of 20th century Scotland, its method and size of our study have given a better estimate than previously available of the heritability of lifespan in modern western populations.

The suggestive curvature of the regression in Figure 2 raises the intriguing possibility that especial long-livedness, say beyond age 90, is more heritable than variation within the normal range and thus survival may not have homogeneous genetic risk basis across all ages. This in turn provides support for GWAS case-control studies of non-agenarians [39], albeit recognising such studies may be focused on the distinct trait of particular livedness.

Returning to overall heritability, Zaitlen et al [18] considered in detail different genomic methodologies for measuring heritability for a range of traits. Their results, which looked at relatedness other than twins, always showed heritability lower than published twin studies. Our results are consistent with this. Zaitlen et al [18] also tried to discern the relative effects of dominance, epistasis and common familial environment, by using different relationships. They showed that analysis based on avuncular relationships gave lower estimates of heritability than parental relationships, which were in turn lower than estimates based on siblings and plausibly concluded that shared environment could account for the differences observed, whilst dominance or epistasis alone could not [18].

We found spousal correlations of similar magnitude to parent offspring correlations, suggesting there is intra-familial resemblance of a non-genetic basis. We agree with Zaitlen et al that modelling common environment is complex, and merits further investigation, noting even their avuncular estimates of heritability may therefore be overstated. It is tempting to estimate the non-genetic correlation between parents and offspring as equal to the whole correlation between spouses, and thus simply subtract spousal correlation from total related pair correlation to estimate genetic correlation. However, we are reluctant to do so, both in principle for any complex trait and for a number of reasons pertinent to lifespan. Firstly some spousal correlation could be genetic, due to assortative mating – most plausibly for a trait like height, but potentially for well-being traits that could affect longevity. However, even if spousal correlation is due to shared environment, it is unclear how that sharing compares with parent-offspring sharing. Given the formative nature of the early years of life, parental influences on aspects of a child’s lifestyle may be greater than that of each parent on the other. On the other hand, assortative mating based on lifestyle choices or perceived well-being, as well as a greater proportion of lifetime spent together, could mean spousal common environment-induced correlations could be greater than those between parents and children. On balance, we believe that it is very likely that there is a common environment induced correlation between parents and offspring in our study, but we have little evidence of its extent. Our heritability estimates therefore represent upper bounds of the heritability of longevity in 20th century Scottish populations.

We found no statistically significant evidence of difference between the sexes in the inheritance of longevity. Nonetheless, the substantially greater mother-daughter lifespan correlation than other parent-offspring correlations in the EASTER study is visually striking. The pattern across the different parent–offspring sex-pairs is similar both in ORCADES and the study of the Old Order Amish by Mitchell et al. [30]. On the other hand, a study of European Royal lineages found much stronger correlation between paternal age-at-death and offspring age-at-death than for mothers and offspring [40] and a study of 18-20th century French Jura Department suggested stronger inheritance of longevity by daughters rather than sons, but with little distinction by the sex of parent [41]. More surprisingly, one study, of unclear size and unusually large heritability estimates, found greatest inheritance across rather than within the sexes [42]. There is thus, as yet, little consensus, on the existence, let alone nature, of sex-based differences in the inheritance of longevity and our study only complicates this picture.

There is a wide range (0.0 – 0.5) of previous, commonly cited, estimates for the heritability of longevity [43], [32]. This wide span of estimates appears to us to arise from four main sources. Firstly heritability of any trait may vary from population to population due to different causal alleles segregating and differing environments[6]. Secondly, different phenotypes are measured. Whilst we have favoured the straightforward age at death approach (over a sixty year span between the ages of 43 and 103), other studies have looked at complex models of mortality risk and given some of the highest estimates for heritability (0.35 and 0.5) [31] [32]. Thirdly, different study designs (eg twin as opposed to parent-offspring) will have different biases. Finally, sampling variance remains quite large even for moderately sized studies, such as the study of the Old Order Amish and our study, where the standard error of the heritability estimates was still 0.05 [30] and 0.02 respectively. Given these complications, we suggest future heritability studies always quote clear standard errors on their estimates, following Beeton and Pearson's good example from as far back as 1899 [25]. We also suggest it would be beneficial if, where possible, researchers (also) gave estimates of heritability for the age at death trait rather than or as well as derived traits such as relative risk statistics or case-control measures, due to the simplicity and comparability of age at death and to facilitate meta-analysis.

Our upper bound estimate of heritability of 0.166 contrasts somewhat with the central estimate of Herskind et at – 0.25 – with unstated standard error [19], despite some reasonable similarities between the studies. In both studies, subjects were born around the end of the 19th Century. Both studies were of a general population in developed Western Europe. However, despite the difference in population and socio-economic differences between Scotland and Denmark, the principal difference does appear to be one of study design. Herskind's study compared monozygotic and dizygotic twins and fitted a model of additive inheritance, dominance and common environmental effects, supposing that the relative inheritance between twin categories of such effects followed their expected values in the case of genetic effects and were equal for common environment. Our study assumed that additive variation passed between parents and offspring but other effects did not. With regard to genetics both study designs appear to correctly model the established modes of inheritance but also capture epistasis, although there is limited power to distinguish between dominance and additive variance in the twin design [6]. However confounding environmental effects cause (different) difficulties for both study designs. The assumption that environmental sharing is the same for both classes of twin is hard to test and seems implausible, with any extra environmental sharing for monozygotic twins, being falsely attributed to genetics. The parent-offspring design assumes that both generations were subject to the same environmental factors affecting the trait, which at a time of societal change may well not be true for lifespan, and is indeed evidenced by the differing lifespans amongst the generations. False attribution of paternity will also reduce estimates of heritability. We thus believe the Herskind method is leading to inflation of heritability, consistent with Zaitlen et al's [18] general findings on twin methods. Furthermore, the Danish study found a model with dominance variation and no common family environment effect or additive genetic variance, provided the best fit [19] . We find such a model entirely implausible, both on general considerations and the contradiction with our findings of substantial parent-offspring correlation that can arise from common family environment effect or additive genetic variance, but not dominance. To be fair, Herskind found an additive model gave almost as good a fit, but their results still estimated no common environmental effects. At the same time, inter-generational changes may have reduced our estimates of heritability, nonetheless, we conclude that our upper bound 0.166 for heritability of longevity is a more persuasive result, with larger studies using further study designs being justified.

Mitchell et al's study of the Amish also estimated the narrow-sense heritability of age-at-death as 0.25, although this time a standard error (0.05) is given. The Amish study uses the same method as ours, parent offspring regression. However this time the population has a very different character – in particular the communal living environment of the Amish. The presumed absence of a family specific shared environment, confirmed by the observed lack of spousal correlations, suggest that Mitchell's study is a particularly sound study of narrow sense heritability. However, it is also plausible that the uniform environment of the Amish is reducing the environmental variance component of longevity and so increasing the proportion of phenotypic variation explained by genetics – and thus narrow sense heritability of lifespan could well be higher in the Amish than populations with a more conventional Western lifestyle. We thus conclude that our upper bound estimate of 0.166 for heritability of lifespan is more relevant to 20th Century populations living a conventional Western lifestyle, than Mitchell's excellent study of a different population.

Our estimate of heritability of longevity (0.16), contrasts with higher estimates of the heritability of the incidence of common killer diseases such as coronary artery disease (0.49) [44] and Alzheimer's disease (0.58) [45], although heritability estimates for cancers are somewhat lower and vary by type [46]. A naïve estimate of the heritability of lifespan would simply be a weighted average of the heritability of the diseases causing death. Our estimate appears lower than that. This could be for a number of reasons. Firstly, for the reasons already outlined, methodological bias may have led to inflation of the estimates of disease heritability. Secondly, there may be less genetic correlation amongst diseases than the environmental correlations (e.g. due to smoking), resulting in a lower genetic proportion for a combined trait. More intriguingly, there may be antagonistic pleiotropy between the disease traits, inducing negative correlations in genetically caused disease susceptibility. Finally, on top of susceptibility to killer diseases, there may be other environmental drivers of lifespan, perhaps affecting frailty, that reduce heritability of lifespan. Indeed, all of these effects may together contribute to the lower heritability of lifespan, than many common diseases, we observed.

The low heritability of lifespan, and the lack of success of existing methods in genetically predicting even complex traits that are more heritable in humans [47], suggests that DNA is unlikely to be an effective predictor of relative lifespan. Instead we postulate that prediction based on biomarkers observed later in life [48], perhaps supplemented by DNA information, will be of more use to individuals and actuaries wishing to predict individual lifespan.

Our study has suggested a revision downwards from commonly cited estimates of the heritability of lifespan in modern western populations to 0.16 or less, perhaps confirming previous doubts over estimates obtained from twin studies. We have found an intriguing, but not statistically significant suggestion of sex-based differences in parent-offspring correlations of longevity, but substantially more data is needed to resolve both the existence and cause of any such differences.

2.5 Supplementary Tables

2.5.1 Analysis of effect of including available Socio-Economic factors in EASTER model.

Method

Occupations were coded according to industry and level. Industry was divided into seven broad categories: mines and quarries (mines); mills and factories (mills); agriculture, horticulture, and animal husbandry or slaughtering (agriculture); transport, delivery, local council labour, and domestic service (service); tradesmen, artists, merchants, and restaurateurs (trades); professional fields such as health, finance, or education (professionals), military and maritime (military). There were also separate categories for wealthy landowners (wealthy) and people whose industries could not be determined (unknown). These categories were based on the Historical International Standard Classification of Occupations (HISCO) (van Leeuwen & Maas 2010), though not all occupations in the dataset were represented in the HISCO database, and in the interest of keeping the number of industry categories down while still leaving room for the "wealthy" and "unknown" groups, not all of the HISCO categories were used.

Level was determined by what job the person performed within a given industry. Level 1 was for generally unskilled or physical labour. Level 2 was for skilled physical labour or supervisors of unskilled workers. Level 3 was assigned to non-physical labour, such as sales or clerical work. Level 4 was for ownership, leadership or strictly managerial positions. These were based on the HISCLASS (van Leeuwen & Maas 2005) categorization method for HISCO (van Leeuwen & Maas 2010). People whose industries were categorized as "unknown" were coded initially as being Level 1, as most of them were "general labourers". If keywords pertaining to other levels appeared in the listed occupation, then they were recoded accordingly.

Age at death residuals were calculated from

Offspring_age_at_death ~ Offspring _location+ husband_level+ husband_industry, and

parent_age_at_death ~ father_level+ father_industry+ offspring_location, As separate location information was not available for parents, and neither was mother-specific occupational details.

Results

Table 4 Analysis of the effect to estimated correlations by fitting available socio-economic variables in the EASTER study

Study	**Offspring**	**Parent**	**Proportion of Variance Explained by covariates Offspring**	**Proportion of Variance Explained by covariates Parents**	**Number of Duos**	**Correlation in ages at death without fitting covariates**	**Correlation in age at death residuals**	**Standard Error of Correlation of residuals**
	Daughter	Mother	0.0318	0.019257	1354	0.1604	0.1513	0.0269
	Son	Mother	0.0281	0.0258447	1630	0.0902	0.0712	0.0247
EASTER	Son	Father	0.0281	0.0152424	1630	0.0659	0.0536	0.0247
	Daughter	Father	0.0318	0.0116785	1354	0.0696	0.0639	0.0271
	Offspring	**Parent**	**0.082**	**0.0152675**	**5968**	**0.0959(1)**	**0.084**	**0.0129**

1: For the multi-sex offspring parent line, the correlation is in age at death residuals having fitted sex as a fixed effect (i.e. age at death ~ sex) to each generation. The proportion of variance in age at death explained by sex for the (multi-sex) aggregate offspring-parent lines, only was 0.0592 (offspring) 0.0036 (parents)

The correlations between parents' ages at death residuals was 0.0680 (compared with 0.0751 in the raw ages at death themselves), n=2,984.

2.5.2 Analysis of Effect of excluding any trio where one member had an extreme age at death

Table 5 Comparison of correlation in Parent and Offspring ages at death, including and excluding extreme ages at death

Study	**Offspring**	**Parent**	**Number of Duos**	**Correlation in ages at death ages >42**	**Correlation in ages at death 50-99 only**	**Standard Error of Correlation ages 50-99 only**
	Daughter	Mother	1145	0.1604	0.1643	0.0292
	Son	Mother	1389	0.0902	0.0773	0.0268
EASTER	Son	Father	1389	0.0659	0.0584	0.0268
	Daughter	Father	1145	0.0696	0.0777	0.0295
	Offspring	**Parent**	**5068**	0.0959	**0.0926**	**0.014**
	Daughter	Mother	665	0.0787	0.0173	0.0388
	Son	Mother	721	0.0697	0.0969	0.0371
ORCADES	Son	Father	721	0.0755	0.1135	0.0371
	Daughter	Father	665	0.0163	0.062	0.0388
	Offspring	**Parent**	**2772**	0.0601	**0.0736**	**0.0189**

Chapter 3 Optimisation of Runs of Homozygosity calling for use in meta-studies using PLINK with a variety of genotyping arrays

3.1 Introduction

Geneticists now commonly use meta-analysis across genetic cohorts to increase statistical power to detect associations between complex traits and genetic factors [2]. As the number of genotyping platforms used by geneticists continue to increase, it becomes more and more desirable to admit a variety of genotyping arrays into such meta-analyses. The meta-analysis technique is particularly well established for GWAS, where imputation is used to create, by panel augmentation, a common comparable set of variants, despite starting with different variants from different arrays [49], However for studies of long runs of homozygosity (ROH), imputation techniques are complex and less well established. Indeed one previous large ROH association meta-study limited participation to one genotyping array, thus avoiding the issue of how to combine arrays [21], although another did use imputation techniques to combine varied arrays [50].

Before proceeding to undertake trait-ROH association studies using multi-array meta-analysis, it is therefore desirable to investigate the effect of genotyping array on measured SROH (summed length of ROH for an individual) using directly typed SNPs and develop a practical protocol which maximises accuracy whilst minimising differences between arrays. The approach has intuitive appeal – ROH should be evident from runs of consecutive SNPs apparent whatever genotyping platform is used and avoids the practical complexities of imputation. PLINK [51] offers a reasonably reliable [52] sliding window based measurement of SROH, which is straightforward to run, whilst offering the possibility of tuning a number of SNP-related parameters [51] thus making it a natural candidate for measuring SROH in meta-analysis. However, assessments of array dependent accuracy and consistency can only be derived from real data if ROH have been accurately measured for the subjects and if the subjects have been genotyped on a variety of arrays.

The 1000 Genomes (1kG) [53] and International HapMap consortium (HapMap) [54] projects have genotyped 851 subjects on three different genotyping arrays, which when combined provide a very dense (2.5M) SNP panel, and, as we shall show, cover the great majority of SNPs on three frequently used commercial arrays. We therefore compared measured SROH from the combined 2.5M panel and three commercial array SNP panels for these subjects, to determine if a PLINK protocol could be developed that was sufficiently accurate, yet relatively insensitive to genotyping array chosen.

3.2 Method

3.2.1 Genotypes

Omni2.5 array genotypes for 2,141 subjects were downloaded from the 1kG website (http://www.1000genomes.org). Similarly, genotypes for 1,184 subjects were downloaded from HapMap (http://hapmap.ncbi.nlm.nih.gov/) for Illumina Human1M and the Affymetrix SNP 6.0 arrays combined.

The HapMap data had already been subject to quality control (QC). The individual population files were merged, while subjects that were not also on the 1kG panel were removed. This gave a HapMap two array dataset with 851 samples genotyped at 1.65M SNPs.

For the 1kG array data, genotypes with a GenCall score less than 0.6 were set to missing. SNPs that were then missing in more than 5% of samples or with a minor allele frequency of less than 3% were removed, whilst samples with more than 3% of SNPs missing were also removed. Again, only samples that appeared on the HapMap panel were retained, giving 851 samples with genotypes at 1.60m SNPs.

The two resultant panels were merged giving a post-quality controlled combined dataset of 851 samples at 2.70m SNPs.

SNP lists for the following commonly used arrays were obtained from http://www.well.ox.ac.uk/~wrayner/strand/: Illumina HAP370CNV (HAP370), HumanOmniExpress-12v1 (OmniX) and Affymetrix GenomeWideSNP 6 (Affy6).

Simulated arrays for the above genotyping arrays were then obtained by extracting the SNP list concerned from the combined panel of 2.7m SNPs. We thus in effect had 851 subjects genotyped on 3 different commonly used arrays and could also look at the particularly dense full panel (ALL).

It is common to prune SNPs for LD prior to ROH calling (ROH-LD), to reduce the preponderance of population wide haplotypes, associated with regions of high LD [36]. After applying the QC and extracting a representation of the three different commercial arrays from the ALL panel, the remaining SNPs were then pruned for pairwise linkage disequilibrium, to remove SNPs within a 50 SNP window that had r2 > 0.1 using the following PLINK options --indep-pairwise 50 5 0.1, consistent with our previous approach [21]. We now had reconstructed unpruned and pruned genotypes for 3 commercial SNP arrays, allowing us to compare and contrast the effect of the array under different ROH calling parameters.

Table 6 shows the numbers of post QC SNPs recovered and the number of autosomal SNPs further cleaned for MAF > 0.05 and genotyping rate >0.03, for each of the 3 genotyping arrays under analysis along with reference count of SNPs sourced for each array. The drop in SNPs from the full list of SNPs on the array was always less than 10%, which looks reasonable, allowing for SNPs dropped in QC, and so the SNP subsets recovered were judged as reasonable proxies for the SNPs that would be genotyped by these chips in a study's post-QC dataset. More importantly, we had 3 SNP panels of differing sizes, from two different array manufacturers, enabling measurement of the effect of using different arrays to determine SROH and SROH-LD.

Table 6 Counts of array SNPs recovered from merged array panel and comparison with count of SNPs in reference list for each array.

Array	Reference Count of SNPs on array	Count of (post QC) SNPs recovered	Autosomal SNPs MAF >0.05 Geno >0.03	Autosomal SNPs MAF >0.05 Geno >0.03 after LD pruning
HAP370	370303	343526	289410	54296
Affy6	934969	907686	569386	61458
OmniX	731345	709756	623795	70801
ALL		2611121	1954192	104744

To simplify visual comparisons individual 1kG populations were amalgamated into continental groupings, in accordance with the usual 1kG nomenclature: AMR (Mixed Americans), EUR (Europeans), ASN (East Asians), SAN (South Asians), AFR (Africans)

3.2.2 ROH Determination

ROH were determined using PLINK [51]. In order to assess the effect of using different genotyping arrays to determine ROH and to develop a protocol to minimise the effect of array, we first sought to develop a gold standard against which array results could be measured. We reasoned that the densest panel available should be able to capture ROH most precisely, providing that effective allowance for genotyping errors, which might falsely break a true ROH, was made.

The PLINK parameters all considered ROH of 1.5Mb and above, but varied the number of SNPs (25-100) needed to form an ROH and the maximum number of permitted (presumed false) heterozygotes (1-2) found in an ROH. The full list of parameters is shown in Table 7.

Table 7 PLINK Parameters considered to create gold standard ROH calling from ALL SNP panel.

Label	SNPs	kb	Gap kb	Density (kb/SNP)	Missing allowed	Hetero-zygotes allowed
SNP:25 Het 1	25	1500	1000	50	5	1
SNP:50 Het 1	50	1500	1000	50	5	1
SNP:50 Het 2	100	1500	1000	50	5	2
SNP:100 Het 2 *	**100**	**1500**	**1000**	**50**	**5**	**2**
SNP:100 Het 3	100	1500	1000	50	5	3

*Parameterisation * using the ALL panel was subsequently adopted as our gold standard.*

--homozyg-window-snp was always set equal to the value of --homozyg-snp

We also used a permitted minimum density (50KB/SNP), consistent with that of the individual array protocol we eventually adopted. The change permits more kb/SNP than McQuillan et al [55] , and, as we show later, is helpful for arrays with sparser coverage. The change in minimum density makes little or no difference for the dense 2.5m SNP panel being used at this stage, but we adopt it here for the sake of consistency with the later analysis.

Our approach was to examine the correlation of measured SROH under the different PLINK parameters. To avoid a few outlying subjects with particularly large SROH dominating the analysis (and improving measured correlations), we considered comparisons with subjects whose SROH did not exceed 30Mb, as well as sometimes including comparisons of all subjects.

3.2.3 ROH-LD Determination

Under LD pruning, care is necessary to ensure that SNP thinning successfully removes population wide haplotypes, even for a particularly dense gold standard reference panel. Bearing this in mind, our analysis proceeded as before.

Again, we examined the sensitivity of the dense panel to different –homozyg parameterisation and then compared the sensitivity of different PLINK parameterisations to the choice of SNP panel. Furthermore, although McQuillan et al [55] had adopted minimum ROH lengths of 1Mb when LD pruned SNPs were considered, we decided to use 1.5Mb to maintain consistency with the unpruned analysis. As for unpruned SNPs, we varied the number of SNPs (12-100) needed to form an ROH (reducing the lower bound to account for the lower number of SNPs after pruning) and the maximum number of (presumed false) heterozygotes (1-2) found in an ROH. The full list of parameters is shown in Table 8.

Table 8 PLINK Parameters considered to create gold standard ROH-LD calling from SNP Panel ALL

Label	SNPs	Kb	Gap kb	Density (kb/SNP)	Missing allowed	Hetero-zygotes allowed
SNP:12 Het 1	12	1500	1000	250	5	1
SNP:25 Het 1	25	1500	1000	250	5	1
SNP:37 Het 2	37	1500	1000	250	5	1
SNP:50 Het 1 *	**50**	**1500**	**1000**	**250**	**5**	**1**
SNP:100 Het 2	100	1500	1000	250	10	2

*Parameterisation * using the ALL panel was subsequently adopted as our gold standard for ROH calling on pruned SNPs.*

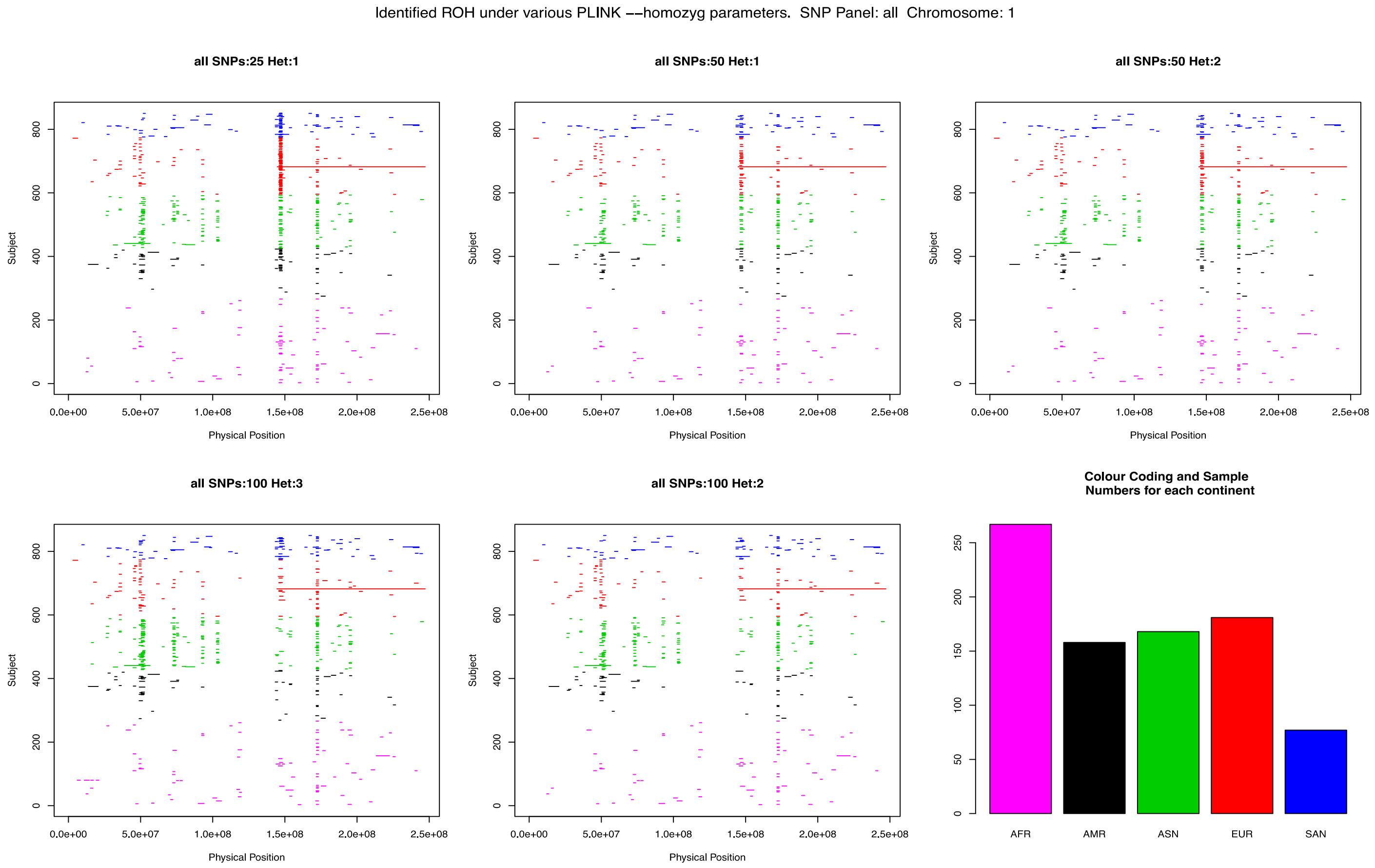
--homozyg-window-snp was always set equal to the value of --homozyg-snp

3.3 Results

3.3.1 Unpruned ROH calling

Our analysis showed that using full SNP panel measured SROH was insensitive to the choice of PLINK options listed in Table 7, as illustrated visually in Figure 4

Figure 4 Comparison of identified ROH (minimum length 1,500kB) for 851 subjects under varying numbers of minimum SNPs and permitted heterozygotes, Chromosome 1, using ALL SNP panel PLINK parameterisations are as defined in Table 7



AFR:African

AMR: American

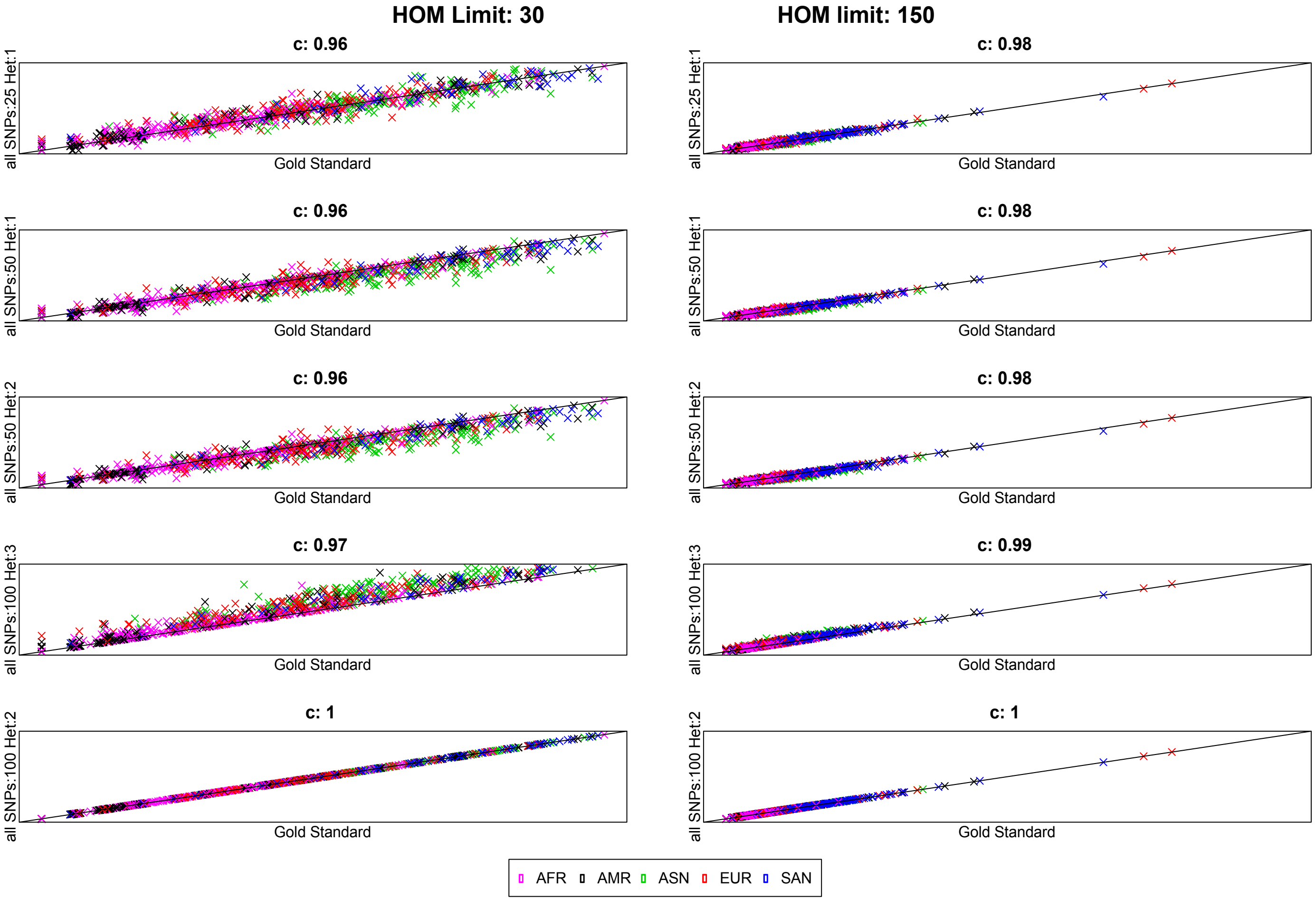
ASN: Asian

EUR: European

SAN: South Asian

The visual similarity in SROH under the different PLINK options considered was confirmed by examination of the correlation of SROH for the ALL panel between the adopted gold standard and the alternative options considered, the regression for which is illustrated in Figure 5. Pairwise correlations between the alternative PLINK commands exceeded 0.97, even when subjects analysed were restricted to the more difficult to call shorter SROH < 30Mb, correlations exceeded 0.94. We also note the different protocols result in slightly different ascertainment levels (i.e. points tending to lie above or below the y=x line, with regression coefficients fitted through zero of 0.91 – 1.14). The effect is limited and in the absence of checking against an external even better gold standard (for example the 1kG sequence data at SNPs latent on the array panels), it was not possible to determine categorically which PLINK parameterisation was objectively best. However, the high degree of concordance meant we felt that further testing was unnecessary. We therefore adopted ROH called using the ALL panel with 100 SNPs and up to 2 heterozygotes in a homozygous window – line SNP:100 Het 2 in Table 7, as our gold standard.

Figure 5 Correlation of measured SROH between gold standard PLINK basis and under alternative PLINK parameters as defined in Table 7



c : correlation

SNP: minimum number of SNPs in a ROH

Het: maximum number of heterozygous SNPs in a ROH

x and y axis limits equal the HOM limit shown.

3.3.2 Pruned ROH-LD calling

We next considered whether a PLINK parameterisation could be determined so as to give a similar close correlation between measured SROH when using the LD pruned SNP panels from different arrays. We reduced the minimum density from 50 kb/SNP to 250 kb/SNP to broadly reflect the five-fold or more reduction in the panel densities, caused by pruning.

The PLINK parameters considered again varied the number of SNPs (12-100) and the number of heterozygotes (1-2) permitted in an ROH, as shown in Table 8, above.

The purpose in LD pruning is to reduce identification of ROH arising from population-wide haplotypes, whilst maintaining identification of ROH arising from more recent consanguinity. Our analysis (illustrated in Figure 6) showed that using the ALL-LD pruned SNP panel, ROH calling was very sensitive to the number of SNPs specified. This is perhaps not surprising – our purpose has been to thin SNPs in regions of high LD to a point where SNPs are sufficiently sparse to be missed by our ROH calling. Sensitivity of ROH calling based on the number of thinned SNPs is thus in essence what we have tried to achieve. Population-wide ROH, for Chromosome 1, can again be identified by the vertical bands in Figure 6. Elimination of these happened as the required number of SNPs rose from 37 to 50. Similar results were obtained for the other chromosomes, except chromosome 6 (Figure 7), where even with a requirement for 100 SNPs, population wide bands are still evident. This result contrasts strongly with that for unpruned SNPs, where measurement was relatively insensitive to the number of SNPs required in a ROH.

Figure 6 Comparison of identified ROH for 851 subjects under varying minimum SNPs and ROH length, Chromosome 1, using ALL SNP panel PLINK parameterisations are as defined in Table 8.

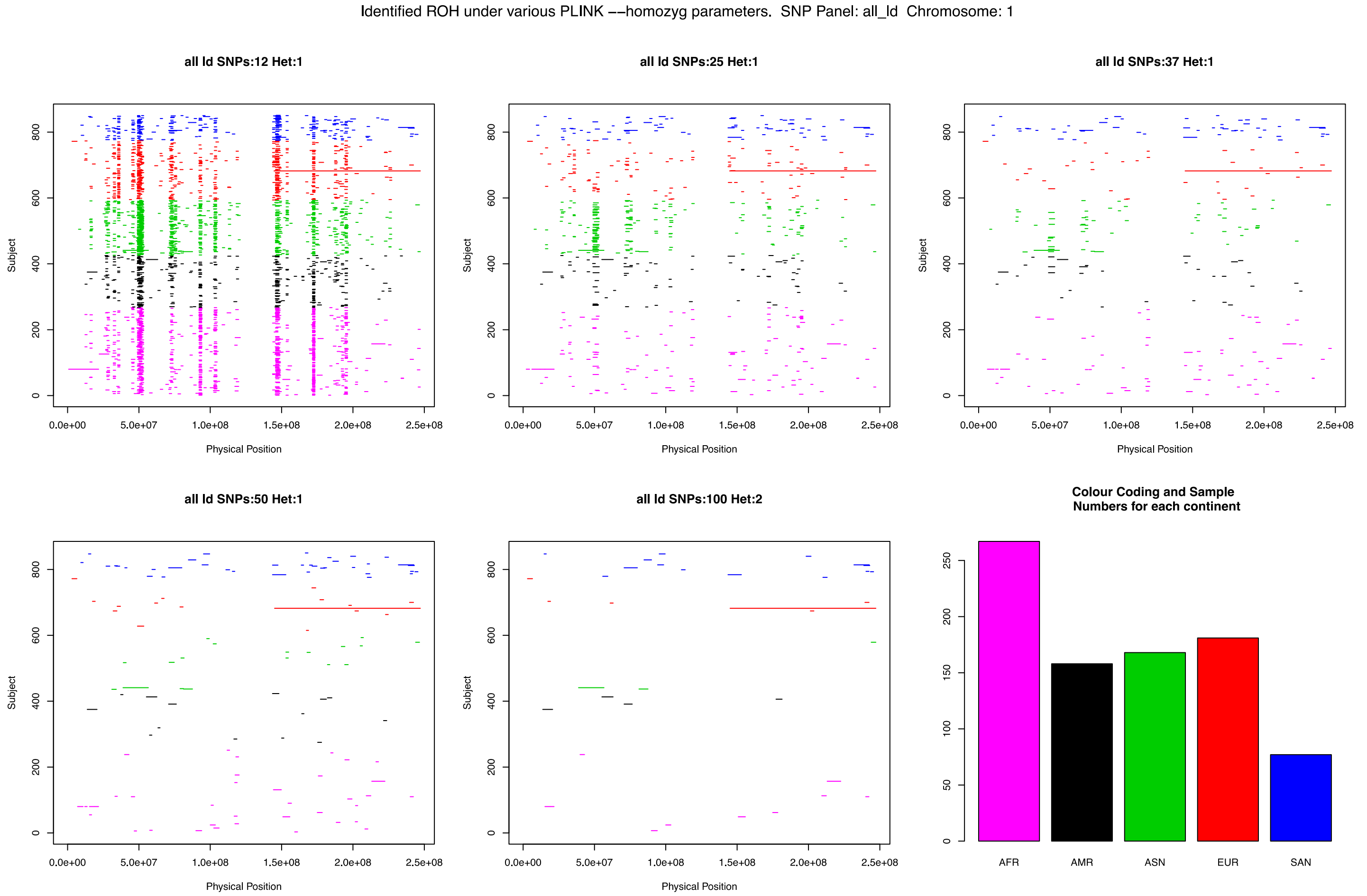


Figure 7 Comparison of identified ROH for 851 subjects under varying minimum SNPs and ROH length, Chromosome 6, using ALL SNP panel PLINK parameterisations are as defined Table 8.

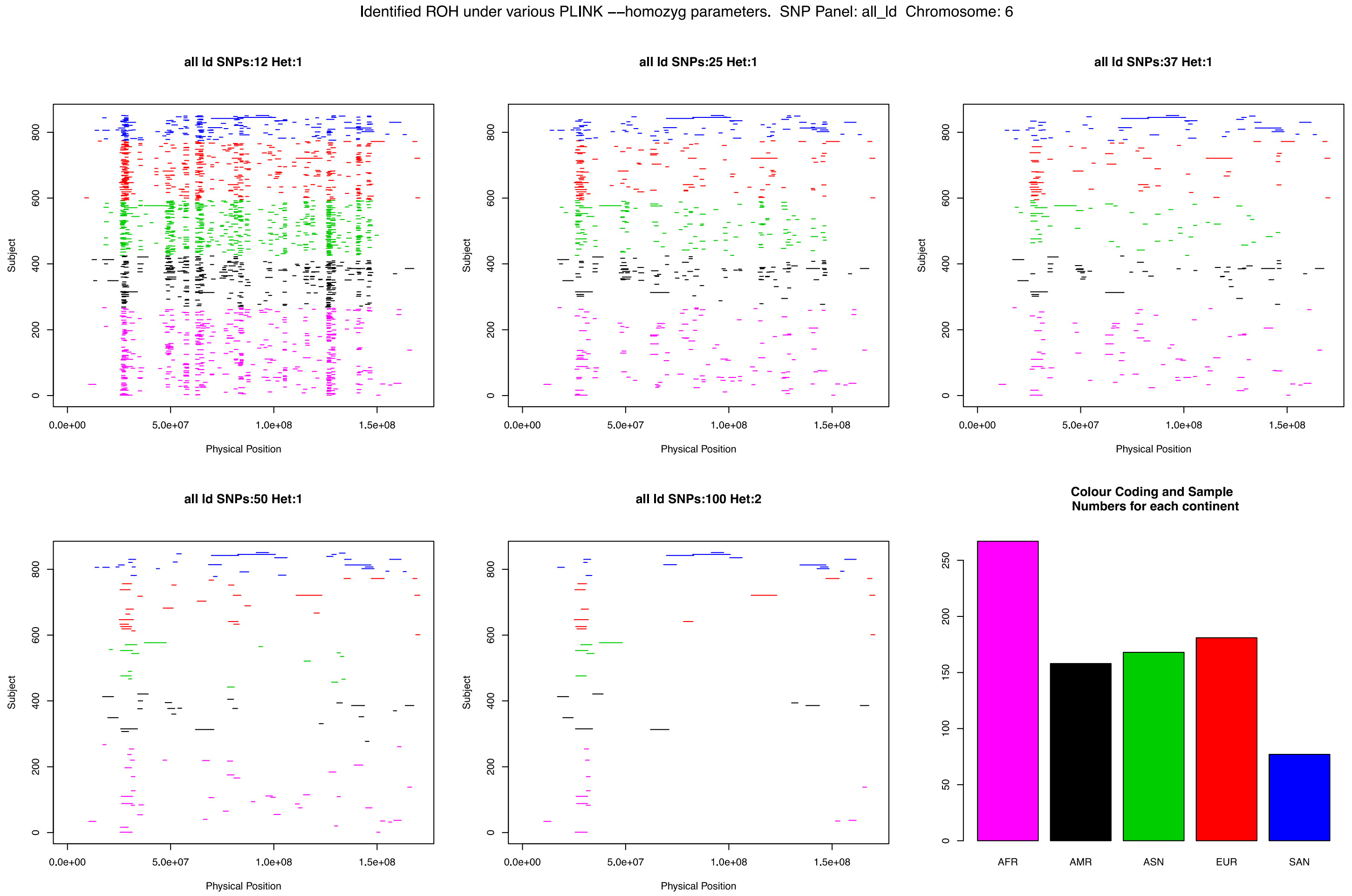
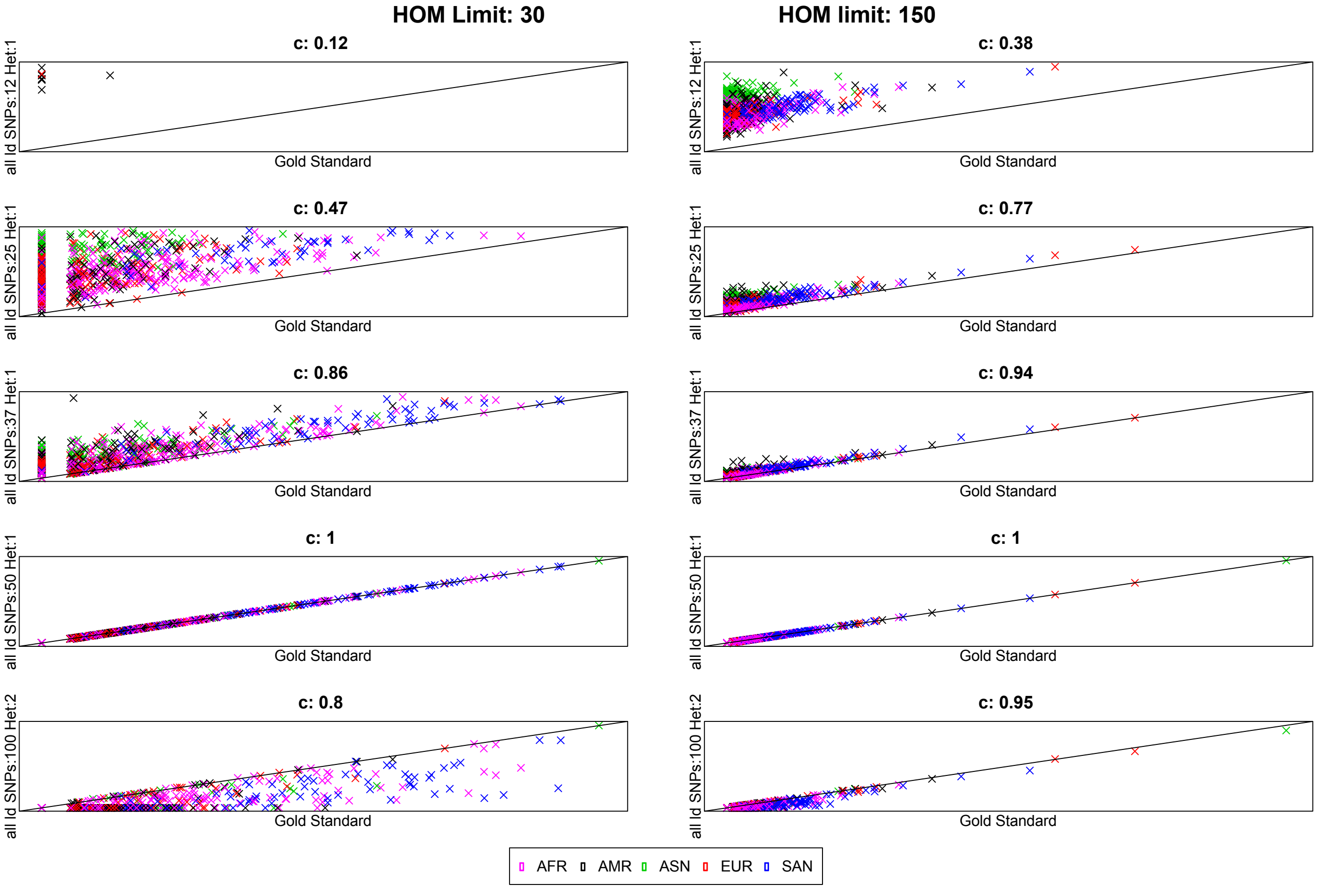


Figure 8 Correlation of measured SROH using ALL-LD Pruned SNPs between gold standard PLINK basis and PLINK parameters as defined in Table 8



c : correlation

SNP: minimum number of SNPs in a ROH

Het: maximum number of heterozygous SNPs in a ROH

x and y axis limits equal the HOM limit shown.

The sensitivity of ROH calling to the number of SNPs required was further confirmed by the low correlation between the different parameterisations. Correlation in SROH between the 12 and 50 SNP requirement was only 0.12, whilst correlation in SROH between the 50 and 100 SNP requirements was 0.8. Furthermore, as was to be expected, as the number of SNPs required increased, less SROH is measured, as illustrated in

Figure 8.

The required numbers of SNPs in an ROH for the ALL panel after LD pruning was thus an, empirically informed, judgment call using the degree of thinning in Figure 4 as a guide, and we settled upon 50 for our gold standard when used with the dense ALL (LD pruned) SNP panel. The need for a judgement call contrasts with the relative insensitivity of results without LD pruning.

Having determined suitable gold standard parameterisations using our ALL very dense SNP panel, before and after LD pruning, we next wished to understand the optimal approach to different panels. In particular, whether ascertainment bias would be substantial between chips and whether a single command could be used for SNP panels used in off-the-shelf genotyping arrays. The PLINK parameters considered varied the number of SNPs required from 25 to 100 and the number of heterozygotes from 1 to 2, as shown in Table 9

.

Table 9 PLINK Parameters considered to call ROH from unpruned SNPs on commercial arrays under consideration

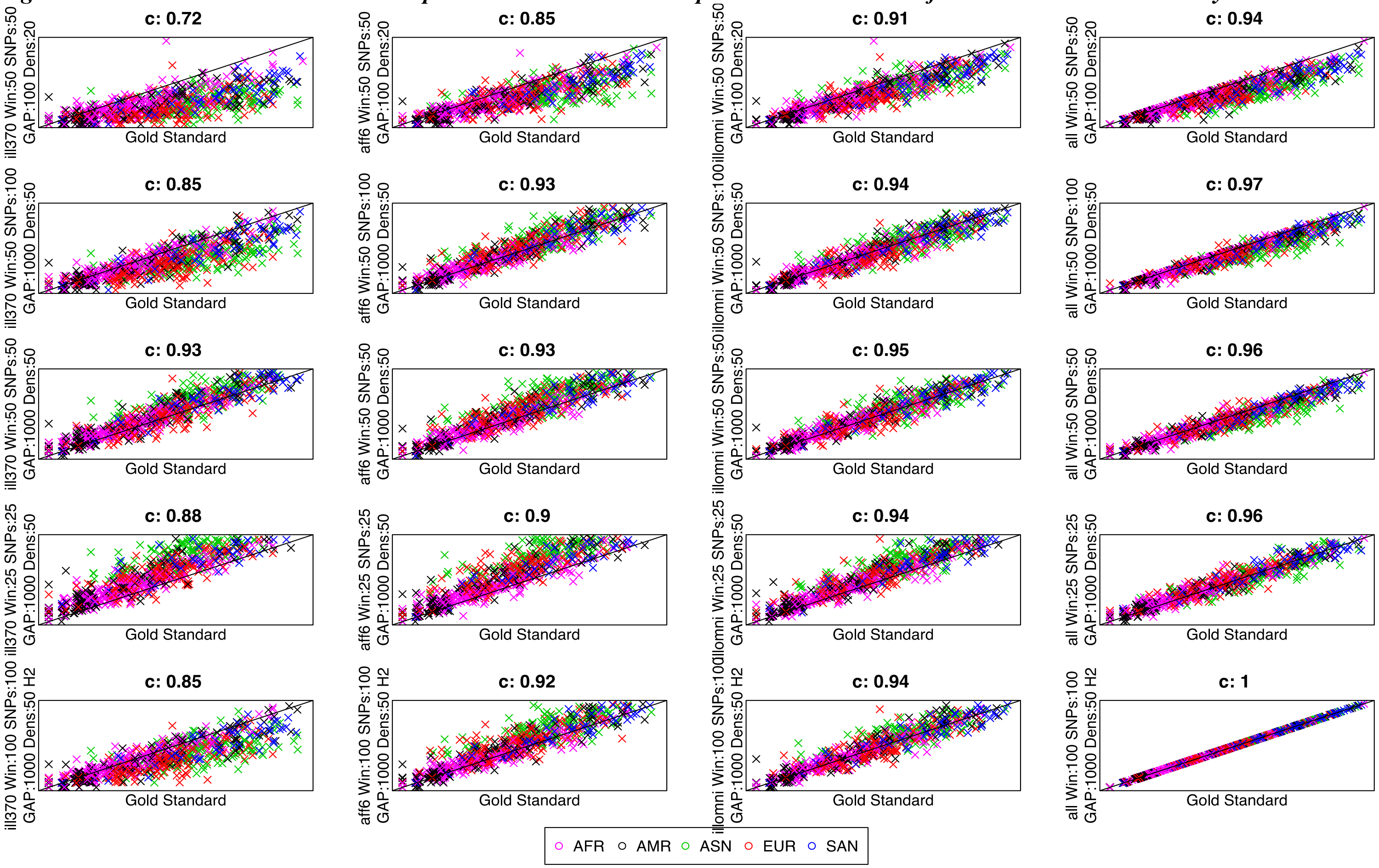
Label	SNP window	SNPs	Kb	Gap kb	Density (kb/SNP)	Missing allowed	Hetero-zygotes allowed
SNP50: Dens20	50	50	1500	100	20	5	1
PLINK-1500	50	100	1500	1000	50	5	1
SNP50	50	50	1500	1000	50	5	1
SNP25	25	25	1500	1000	50	5	1
SNP100	100	100	1500	1000	50	10	2

Requiring 50 SNPs and a density of 20 kb/SNP is close to that adopted by McQuillan et al [21] (although here –snp has been explicitly set to the *de facto* override perhaps unintentionally imposed in McQuillan et al [21] by their –window-snp option). We also consider PLINK defaults, subject to using 1,500 kb minimum length, with other parameterisations varying the number of SNPs needed from the PLINK default parameterisation.

We found that for the less dense HAP370 panel, correlations between the Gold Standard and bases SNP50:Dens20 and PLINK1500 were low (0.68 and 0.82 respectively). Both parameterisations were often underestimating SROH, as Figure 9 shows (points generally fall below the y=x line in rows 1 and 2 of column 1). In both cases areas with sparser SNP coverage were being missed, in PLINK's case due to the insistence of 100 SNPs, whilst for SNP50:Dens20 the density criterion was too stringent. Consistent with this, in both cases SROH estimation improved with the denser arrays (points lie closer to the y=x line in columns 2,3 and four of columns 1 and 2).

On the other hand, the SNP25 and SNP50 parameterisations achieved correlations with the gold standard in excess of 0.88 and 0.93 respectively, for all the panels considered, again with improving correlation as the panel density increased. Furthermore the SNP50 parameterisation does not appear significantly to systematically over or under estimate SROH, relative to the gold standard – points straddle the y=x line in rows 3 of Figure 6 and the regression coefficients are close to, but slightly exceed, 1 (1.09. 1.12 and 1.01 for Illumina HAP300, Affymetrix 6 and Illumina Omni respectively). Results permitting 2 heterozygotes (SNP:100 Het 2) have similar, but slightly worse correlations, and, perhaps not surprisingly, over estimate SROH particularly for the less dense panel – the final row of figure 6.

Figure 9 SROH under Gold standard compared with various PLINK parameterisations and for various commercial arrays



c: correlation

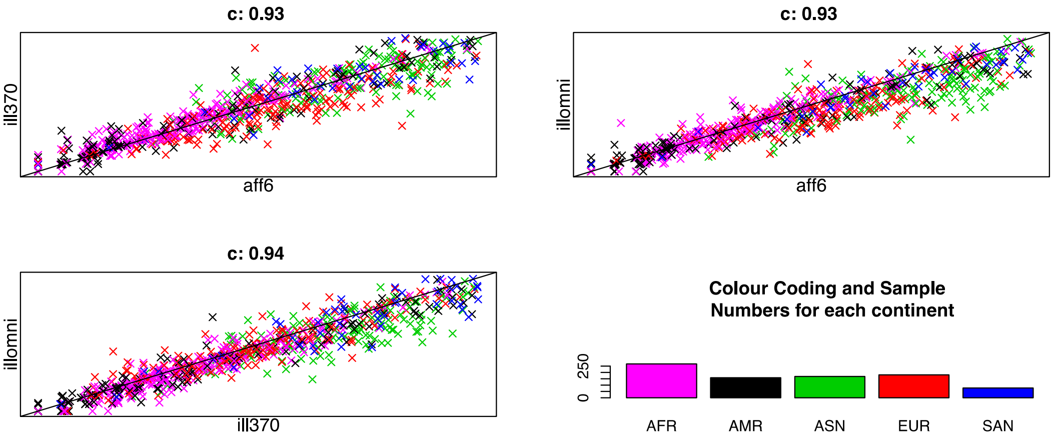
PLINK parameterisations are as defined in Table 9

x and y axis limits are [0,30Mb].

These results show that for 1500kb ROH, the PLINK default parameterisations, except that the minimum SNPs be reduced from 100 to 50, provide closest correspondence to the gold standard determined using a dense chip and suggest that this is a good basis for estimating true SROH across varied unpruned SNP panels, with typical density available on off-the-shelf arrays.

Whilst accuracy of called SROH is perhaps the most important feature of a protocol, we were also interested in the extent of ascertainment bias that might be present. I.e. whether although all reasonably accurate, SROH called from different SNP arrays might differ between the arrays. Admittedly if called SROH from two arrays is close to the truth, the difference between them cannot be that great, but we still felt it important to measure the difference. We found that pair-wise correlations between the three commercial arrays considered, using the SNP:50 protocol, exceeded 0.93, as illustrated in Figure 10.

Figure 10 SROH across different genotyping platforms under adopted PLINK homozyg parameters



c : correlation

x and y axis limits are [0,30Mb].

The slopes of the regression, fitted through the origin, in Figure 10 are 0.959, 0.886, and 0.914 for Aff6-Ill370, Aff6-IllOmni and Ill370-IllOmni respectively, with standard errors in the slope estimates less than 0.005.

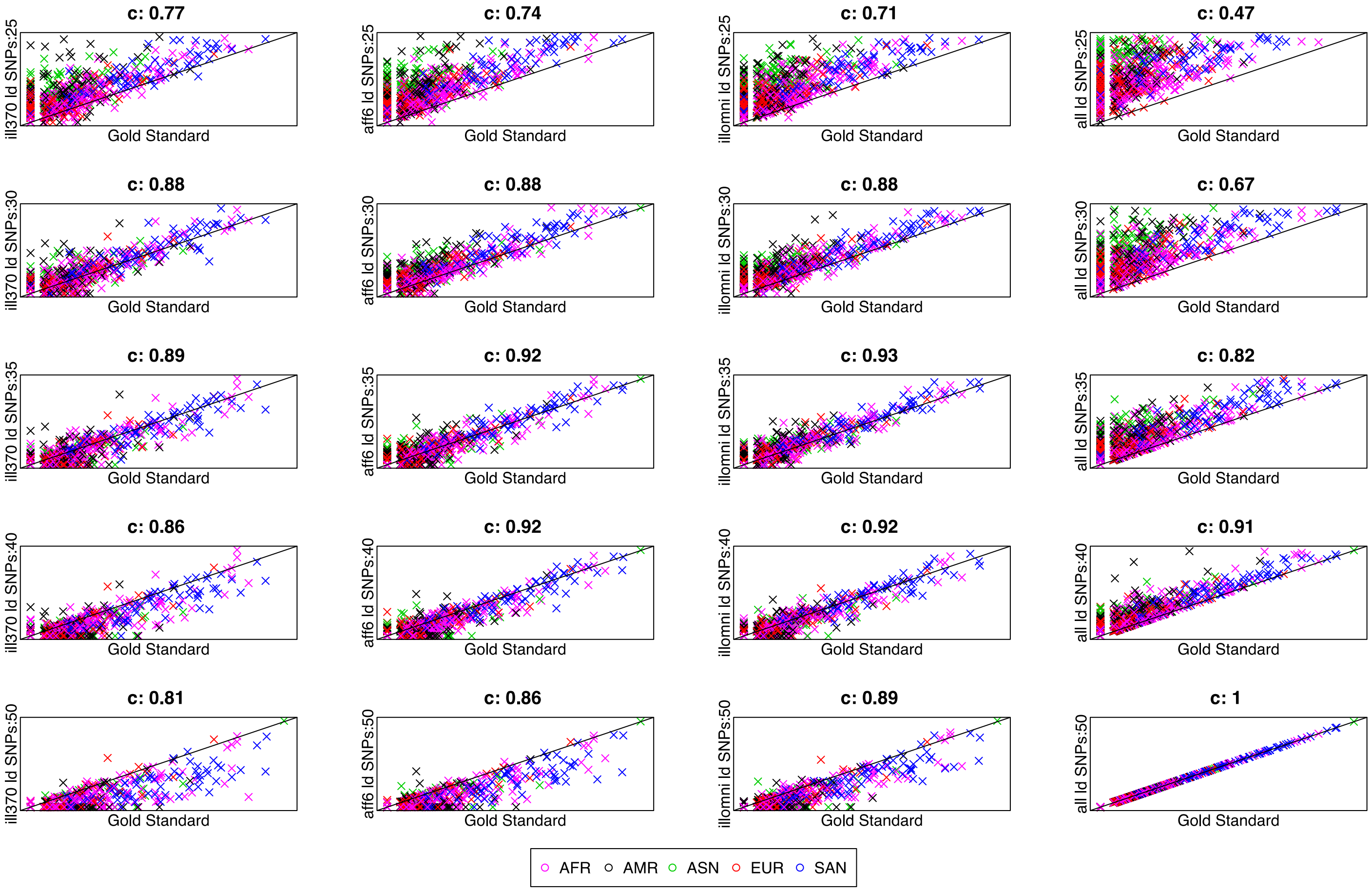
We repeated the protocol optimisation for the LD-pruned panels and considered the number of SNPs required in an ROH in the range of 25-50, whilst permitting a gap of 1Mb between ROH for the ROH to be combined and requiring density to exceed 250kb/SNP, as shown in Table 10.

Table 10 PLINK Parameters considered to call ROH from pruned SNPs on commercial arrays under consideration

Label	SNPs	Kb	Gap kb	Density (kb/SNP)	Missing allowed	Heterozygotes allowed
SNP25	25	1500	1000	250	5	1
SNP30	30	1500	1000	250	5	1
SNP35	35	1500	1000	250	5	1
SNP40	40	1500	1000	250	5	1
SNP50	50	1500	1000	250	5	1

For the commercial array panels and the range of numbers of SNPs considered (25-50) correlations exceeded 0.7, and were sensitive to the numbers of SNPs specified. For all three panels, results were most highly correlated with the gold standard when 35 SNPs where required in a ROH (correlation 0.89, 0.92, 0.93 for HAP370, Affy6 and OmniX respectively), as illustrated in Figure 11,which also shows correlations with the ALL panel, which would not be available in practice. Furthermore the SNP35 parameterisation does not appear significantly to systematically over or under estimate SROH – points straddle the y=x line in rows 3 of Figure 11 - and the regression coefficients are close to 1 (0.877, 0.993 and 1.07 for Illumina HAP300, Affymetrix 6 and Illumina Omni respectively).

Figure 11 SROH-LD under Gold standard compared with various PLINK parameterisations and for various commercial arrays

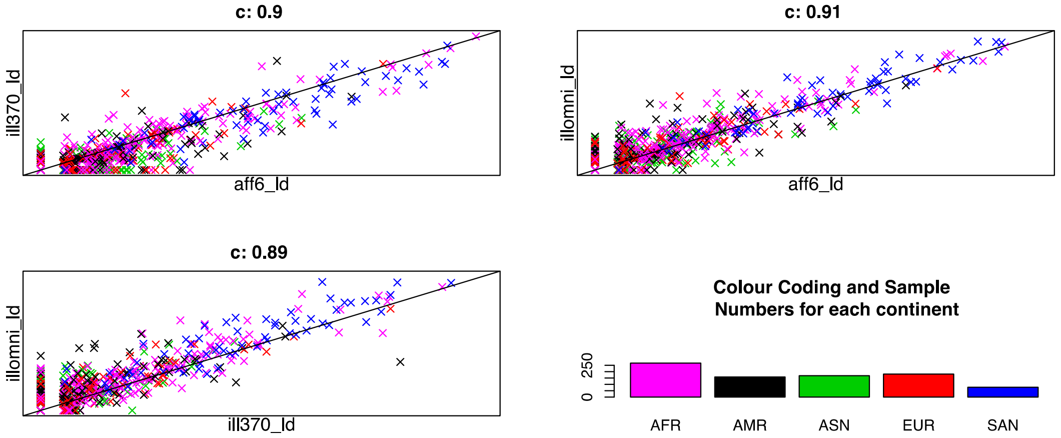


c: correlation

PLINK parameterisations are as defined in Table 10

x and y axis limits are [0,30Mb].

Figure 12 SROH-LD across different genotyping platforms under adopted PLINK homozyg parameters



c : correlation

x and y axis limits are [0,30Mb]

Finally, we again wished to examine SROH-LD correlations between the different arrays. We found pairwise correlations between the chips of 0.89-0.91 comparison was restricted from subjects with SROH-LD < 30000 , as shown in Figure 12 whilst under the easier test, when all subjects were included, correlation always exceeded 0.98. The slopes of the regression, fitted through the origin, in Figure 12 are 0.84, 1.02, and 0.89 for Aff6-Ill370, Aff6-IllOmni and Ill370-IllOmni respectively, with standard errors in the slope estimates less than 0.005.

Thus our analyses suggests that the adopted PLINK commands are suitable for the determination of SROH and SROH-LD, for commercially available SNP panels with SNP counts in the range we considered.

3.4 Discussion

Our newly adopted parameterisation for determining SROH is the PLINK default, except the minimum ROH length is 1,500 kb (vs. 1,000 kb as PLINK default) and 50 SNPs (vs.100 as PLINK default) are required to call a ROH. Whilst ROH length is predominantly a matter of preference for recent or more distant consanguinity detection, the number of SNPs required is important and our adopted number of SNPs is more appropriate for typical genotyping arrays with 300k-800k post-QC SNPs, even for our longer ROH and denser unpruned panel. For determining SROH-LD, our preferred parameterisation again uses the PLINK defaults except that the minimum ROH length is 1,500 kb, the required SNP density is a least one SNP per 250kb (vs 50kb as PLINK default) and 35 SNPS (vs 100 as PLINK default) are required to call a ROH. Once more, other than ROH length, our adopted parameters appear more appropriate than PLINK defaults, which were possibly not designed for such heavily pruned panels.

These parameterisations accord closely with the approach of McQuillan et al [21], which used the default requirement for a 50 SNP sliding window, thus broadly, and presumably inadvertently, at least 50 SNPs in a ROH, despite having set the explicit SNP requirement as 12 under LD pruning. Our reduction in density allows areas of low coverage to be captured and increases correlation with our denser gold standard, panel suggesting that the capture of these areas is not creating false positives.

Although our pruning strategy differed, our results suggesting 35-50 SNPs as being optimal appear somewhat consistent with those in a recent study by Howrigan *et al* in optimisation of IBD calling using simulated data[52], which found that 35-65 SNPs were appropriate. However, closer examination suggests that the differences between that study and this study here, eg with respect to ROH length, the form of gold standard, and the much lighter pruning in Howrigan *et al*'s study, all mean it would be easy to over-interpret the significance of the agreement.

Our results unambiguously suggest that our preferred parameterisation for measuring SROH and SROH-LD is a suitable for use in a multi-array ROH consortium, given the practical and accuracy considerations. The protocol works well across the array densities typically used by genetic consortium members. The protocol is simple and independent of array. At the same time, measured SROH and SROH-LD are closely correlated with our best available estimate of reality and measured SROH and SROH-LD are closely correlated across genotyping arrays. Although measured SROH is sensitive to the genotyping array used, the sensitivity is limited and will still permit effective detection of the effect of SROH on traits using meta-analysis across genotyping arrays, should it exist. We therefore recommend the following PLINK parameterisations for ROH consortia measuring 1.5Mb ROH

SROH: --homozyg --homozyg-window-snp 50 --homozyg-snp 50 --homozyg-kb 1500 --homozyg-gap 1000 --homozyg-density 50 --homozyg-window-missing 5 --homozyg-window-het 1; and

SROH-LD: --homozyg --homozyg-window-snp 35 --homozyg-snp 35 --homozyg-kb 1500 --homozyg-gap 1000 --homozyg-density 250 --homozyg-window-missing 5 --homozyg-window-het 1.

This is, of course, the parameterisation we shall adopt in Chapter 4.

Chapter 4 The effect of genome-wide homozygosity on 16 complex traits

Abstract

Cousin marriage has long been associated with rare, often devastating, Mendelian disorders [56] and Darwin was one of the first to recognise that such inbreeding reduces evolutionary fitness in plants [57] . However, the effect of the more distant parental relatedness common in modern human populations is less well understood. Genomic data now allow us to investigate the inbreeding effects on traits of public health importance by measuring the overall length of homozygous segments (runs of homozygosity, ROH), which are inferred to be inherited identical-by-descent from a common ancestor. Given the low levels of inbreeding prevalent in most human populations, information is required on very large numbers of people to provide sufficient power [21,36]. Here we use ROH which reflect both recent and remote inbreeding in a study of 16 health-related quantitative traits in up to 354,224 individuals from 102 cohorts and find statistically highly significant associations between individual genome-wide summed runs of homozygosity (SROH) and four complex traits: height, forced expiratory lung volume in 1 second (FEV1), general cognitive ability (*g*) and educational attainment (nominal $p<1 \times 10^{-300}$, 2.1×10^{-6}, 2.5×10^{-10}, 1.8×10^{-10}). In each case increased homozygosity was associated with decreased trait value. Similar effect sizes were found across four continental groups and in populations with markedly different degrees of mean inbreeding, providing convincing evidence for the first time that homozygosity, rather than other genetic or environmental confounding, contributes to observed phenotypic variance. Contrary to earlier reports in substantially smaller samples [58,59], no evidence was seen of an influence of background inbreeding on blood pressure and low density lipoprotein (LDL) cholesterol, or ten other cardio-metabolic traits. Since inbreeding depression is predicted for traits under directional evolutionary selection[20] , this study

provides evidence that increased stature and cognitive function have been positively selected in human evolution, whereas many important risk factors for late-onset complex diseases have not.

4.1 Introduction

The inheritance of complex traits has non-additive components as well as additive ones. This is most obvious for Mendelian traits, although for complex traits the extent of dominance variance is difficult to measure, due to environmental confounding [6] . A special form of non-additivity, known as directional dominance, arises when the dominance is biased in one direction on average over all causal loci, for instance to decrease the trait, rather than dominant alleles at some loci increasing the trait value while others decrease it. Such directional dominance is expected to arise in evolutionary fitness-related traits due to directional selection[60]. Non-additive directional dominance gives rise to a subtle, but potentially interesting form of complex trait inheritance.

Directional dominance can be observed through the study of inbreeding, since inbreeding influences complex traits through increases in homozygosity (especially of minor homozygotes) and corresponding reductions in heterozygosity, most likely resulting from the increased action of minor deleterious (partially) recessive mutations[20][1]. Historically inbreeding has been measured using pedigrees [61,62]. However, such techniques cannot account for the stochastic nature of inheritance, nor are they practical for the capture of the distant parental relatedness present in most modern day populations. High density genome-wide single nucleotide polymorphism (SNP) array data can now be used to assess inbreeding directly, using genomic runs of homozygosity (ROH), as called by genomic analysis software packages. Such runs are inferred to be homozygous-by-descent and are common in human populations[63,64]. An additional method of ROH-calling first prunes the SNP panel for linkage disequilibrium (LD) and thus disregards many of the ROH created by

[1] Note – for complex traits we shall use the words dominance and recessivity almost interchangeably as the (+/-) sign of dominance deviation is arbitrary, depending on the (arbitrary) choice of scale.

common segregating haplotypes (showing strong LD), leaving those comprised of alleles in weak LD, which originate from more recent pedigree loops. SROH, or SROH* (for LD-pruned ROH calling), is the sum of the length of these ROH, in megabases of DNA. A genomic version of Wright's inbreeding coefficient, F_{ROH} (F_{ROH*}), may now be inferred as the ratio of SROH (SROH*) to the total length of the genotyped genome. Like pedigree-based F (with which it is highly correlated[36]), F_{ROH} estimates the excess probability of being homozygous at any site in the genome. F_{ROH} has been shown to vary widely within and between populations[55] and is a powerful method of detecting inbreeding effects[65].

4.2 Method

4.2.1 Summary

We meta-analysed the regressions of traits on SROH for 159 sub-cohorts. Sub-cohorts were created from 102 population-based or case-control genetic studies, by separating different genotyping arrays, cases and controls or ethnic sub-groups to ensure each sub-cohort was homogeneous. The full list of participating sub-cohorts is shown at Appendix 1a ROHgen Participating cohorts). Where a sub-cohort had been ascertained on the basis of a disease status associated with a particular trait, that sub-cohort was excluded from the corresponding trait analysis.

ROH with a length of at least 1.5 Mb were called from quality controlled dense genome-wide SNP array data (minimum 300,000 markers), using PLINK (--homozyg --homozyg-window-snp 50 --homozyg-snp 50 --homozyg-kb 1500 --homozyg-gap 1000 --homozyg-density 50 --homozyg-window-missing 5 --homozyg-window-het 1). In chapter 3, we showed that SROH called with these parameters is relatively insensitive to the density and type of array used. The association between the trait and SROH was measured using a linear model in R; Trait ~ SROH + age + sex. In addition the first three within cohort principal components of the relationship matrix were fitted, as were any other cohort-specific covariates known to be associated with the trait, including further principal components, and any trait-specific covariates such as medication. For family-based studies, we also fitted the mixed model, using grammar+ type residuals and full hierarchical mixed modelling using GenABEL and hglm. In the principal results, effect sizes estimated using hglm, where available, were used over grammar+, which were in turn used in preference over fixed model-only estimates. Inverse-variance meta-analysis of all sub-cohorts' effect estimates was performed using Rmeta[38] , on a fixed effect basis

4.2.2 Cohorts

Data from 102 independent genetic epidemiology studies were included. All subjects gave written informed consent and studies were approved by the relevant research ethics committees. Homogeneous sub-cohorts were created for analysis on the basis of ethnicity, genotyping array or other factors. Where a cohort had multiple ethnicities, sub-cohorts for each separate ethnicity were created and analysed separately. In all cases European-, African-, South or Central Asian-, East Asian- and Hispanic-heritage individuals were separated. In some cases sub-categories such as Ashkenazi Jews were also distinguished. Ethnic outliers were excluded, as were the second of any monozygotic twins and pregnant subjects. For case-control and trait extreme studies, patients or extreme-only sub-cohorts were analysed separately to controls. Where case status was associated with the trait under analysis the sub-cohort was excluded from that study (see below).

Subjects within a sub-cohort were genotyped using the same SNP array, or where two very similar arrays were used (e.g. Illumina OmniExpress and IlluminaOmni1), the intersection of SNPs on both arrays – provided the intersection exceeded 300,000 SNPs. Where a study used two different genotyping arrays, separate subcohorts were created for each array, and analysis was done separately. Paediatric cohorts were not included.

4.2.3 Genotypes.

All subjects were genotyped using high density genome-wide (>300,000 SNP) arrays, from Illumina, Affymetrix or Perlegen. Custom arrays were not included. Each study's usual array-specific genotype quality control standards for genome-wide association were used and are shown in Appendix 1b ROHgen Intra-Cohort Genotype QC).

4.2.4 Phenotypes

We studied 16 quantitative traits which are widely available and represent different domains related to health, morbidity and mortality: height, body mass index (BMI), waist: hip ratio (WHR), diastolic and systolic blood pressure (DBP, SBP), fasting plasma glucose (FPG), fasting insulin (FI), Haemoglobin A1c (HbA1c), total-, HDL- and LDL-cholesterol, triglycerides, forced expiratory volume in 1 second (FEV1), ratio of FEV1 to forced vital capacity (FVC), general cognitive ability (*g*) and years of educational attainment (EA). Phenotypic QC was performed locally to assess the accuracy and distribution of phenotypes and covariates. Further covariates were included when the relevant GWAS consortium also included them. The trait categories were anthropometry, blood pressure, glycaemic traits, classical lipids, lung function, cognitive function and educational attainment, following models in the GIANT [2], ICBP [11], MAGIC[66] , CHARGE[67] , and Spirometa [68] and SSGAC [69] consortia. By happy accident, and unlike Spirometa, the model for FEV1 did not include height as a covariate. Effect sizes for FEV1 therefore include size effects that also underpin height. We considered re-analysis adding height as a covariate, however this was impractical due to the scale of the meta-study, but does mean that the FEV1 study is measuring scale rather than lung function independent of scale, which would be an interesting, entirely separate, albeit interesting, trait. Studies assembled files containing study traits and the following covariates: sex, age, three principal components of ancestry, lipid-lowering medication, ever-smoker status, anti-hypertensive medication, diabetes status and year of birth (YOB). Educational attainment was defined in accordance with the ISCED 1997 classification (UNESCO), leading to seven categories of educational attainment that are internationally comparable [69]. LDL values estimated using Friedewald's equation were accepted. Cohorts without fasting samples did not participate in the LDL-cholesterol, triglycerides, fasting insulin or fasting plasma glucose analyses. Cohorts with semi-fasting samples fitted a categorical or quantitative fasting time variable as a covariate. Subjects with less than 4 hours fasting were not included.

Where subjects were ascertained, for example, on the basis of hypertension, that sub-cohort was excluded from analysis of traits associated with the disorder, for example blood pressure. A list of traits excluded by sub-cohort ascertainment is shown in Table 11

Table 11 Traits where cases were excluded

Ascertainment basis	**Traits for which Cases excluded**
Type-2-Diabetes	Fasting insulin, HbA1c, fasting plasma glucose
Hypertension	Blood pressures
Venous thrombosis, Coronary artery disease (CAD)	Blood lipids
Obesity, metabolic syndrome	As CAD, plus BMI, waist-hip ratio, fasting insulin and fasting plasma glucose

Somewhat unusually for a large consortium meta-analysis, the majority of the analysis after initial genotype and phenotype QC was performed by a pipeline of standardised R and shell scripts, to ensure uniformity and reduce the risk of errors and ambiguities. The standardised pipeline was used for all stages from this point onwards.

4.2.5 Calling Runs of Homozygosity.

As a further layer of QC, pre ROH-calling, SNPs with more than 3% missingness across individuals or with a minor allele frequency less than 5% were removed.

Autosomal runs of homozygosity exceeding 1.5 Mb in length were called using PLINK [51] , with the following settings as specified in Chapter 3--homozyg-window-snp 50 --homozyg-snp 50 --homozyg-kb 1500 --homozyg-gap 1000 --homozyg-density 50 --homozyg-window-missing 5 --homozyg-window-het 1. LD pruned runs of homozygosity were called, after excluding non-independent SNPs with r^2 exceeding 0.1, using the same parameters except --homozyg-window-snp 35 --homozyg-snp 35 --homozyg-density 250. The sum of runs of homozygosity was then calculated for unpruned (SROH) and pruned (SROH*) SNPs[4]. F_{ROH} was calculated as SROH/($3x10^9$) reflecting the length of the typically genotyped autosomal genome. Copy number variants (CNV) are known to influence cognition[38]; however, prior calling of CNV and ROH in one of our cohorts reduced the SROH by only 0.3%[36], making it implausible that deletions called as ROH influence our findings.

4.2.6 Trait association with SROH.

The association between trait and SROH or SROH* was calculated using a linear model

trait ~ SROH (or SROH) + age + sex + pc1 + pc2 + pc3.*

Pc1-pc3 were the first 3 principal components of the genetic relationship matrix. Additional covariates were fitted for some analyses (shown below) or for some cohorts where analysts were aware of study specific effects (e.g. study centre). For BMI, WHR, FEV1, FEV1/FVC and *g*, trait residuals were calculated for the model excluding SROH, these residuals were then rank-normalised and the effect of SROH on these rank-normalised residuals estimated. Triglycerides and fasting insulin were natural log transformed. Additional covariates were as follows: age^2 was included as a covariate for all traits apart from height and *g*. BMI was included as a covariate for

WHR, SBP, DBP, FPG, FI and HbA1c. Year of birth (YOB) was included a covariate for EA and ever-smoking for FEV1 and FEV1/FVC. Where a subject was known to be taking lipid-lowering medication, total cholesterol was adjusted by dividing by 0.8. Similarly, where a subject was known to be taking anti-hypertensive medication, SBP and DBP measurements were increased by 15 and 10 mm Hg, respectively.

4.2.7 Relatedness.

Where the cohort was known to have significant kinship, genetic relatedness was also fitted, using the mixed model. The polygenic model was fitted in GenABEL using the fixed covariates and the genomic relationship matrix[70] . GRAMMAR+ (GR+) [71] residuals were then fitted to SROH as well as the full mixed model being fitted simultaneously, using GenABEL's hierarchical generalised linear model (HGLM) function [72] . Populations with kinship thus potentially had six estimates of β_{FROH}: using fixed effects only, and using the mixed model approaches, GR+ and HGLM) for SROH and SROH*.

4.2.8 Confounding

To investigate potential confounding, where available, EA was added as an ordinal covariate and all models rerun, giving revised estimates of β_{FROH}. This is potentially an over adjustment for *g* due to the phenotypic and genetic correlations with EA[73]. Meta-analyses were rerun for various subsets, according to geographic and demographic features of the cohorts. Cohorts were divided into more homozygous and less homozygous strata with the boundary being set so each within-stratum meta-analysis had equal statistical power (simply by summing the rank ordered weights in the base analysis).

Cohort phenotypic means and standard deviations were checked visually for inter-cohort consistency, with apparent outliers then being corrected (e.g. due to units or incorrectly specified missing values), explained (e.g. due to different population characteristics) or excluded.

4.2.9 Meta-analysis

In the principal analyses, for cohorts with relatedness, HGLM estimates of β_{FROH} were preferred, however where HGLM had failed to converge, results using GR+ were included. These results were combined with those for unrelated cohorts which had been determined on a fixed effect only basis. Result outliers were defined as individual cohort by trait results, which failed the hypothesis, cohort (β_{FROH}) = pre-QC meta-analysis (β_{FROH}), with a t-test statistic >3. Analyses were performed with and without outliers for β_{FROH} in phenotypic units and in intra-sex phenotypic standard deviations. The principal results we present are for F_{ROH} (i.e. no LD pruning), with outliers included for the hypothesis tests (which turns out to be more conservative), but with outliers excluded when estimating β_{FROH} [74] . Meta-analysis was performed using inverse variance meta-analysis in the R package Rmeta [38] , with β_{FROH} taken as a fixed effect and alternatively as a random effect.

4.2.10 Heritability in ORCADES

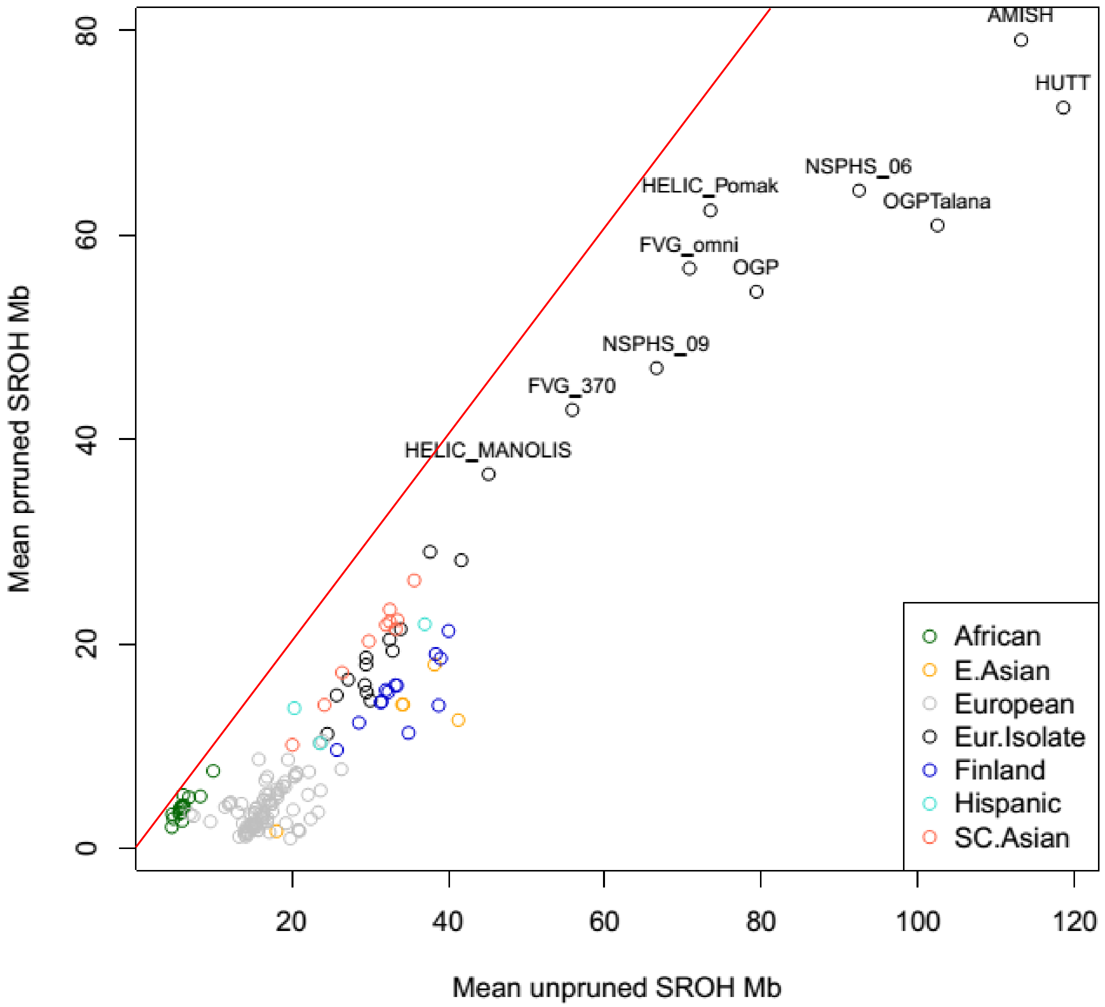
Heritabilities and correlations for the four traits showing ID were calculated in ORCADES using bivariate restricted maximum likelihood in GCTA [17] and the same base phenotypes and genotypes as the main study.

4.3 Results

We found marked differences by geography and demographic history in both the population mean SROH and the relationship between SROH and SROH* Figure 13. As observed previously [36,55], isolated populations have a higher burden of ROH whereas African heritage populations have the least homozygosity. Populations differed by an order of magnitude in their mean ROH. There are clear differences by continent and population type both in the mean SROH, and the relationship between pruned and unpruned ROH. Isolated populations have a higher amount of SROH whereas African heritage populations have the least homozygosity. Africans also have less unpruned SROH for their level of pruned ROH (they are nearer to the y=x line in Figure 13), suggesting that they have fewer shorter ROH and that more of their ROH are of recent pedigree, although inter-continental differences in LD complicate the picture. These results are consistent with the out of Africa hypothesis [75], that humans emerged from Africa to populate other parts of the globe, creating population bottlenecks in non-African populations. This is expected to give rise to more population-wide long haplotypes in non-African populations, and thus more and longer ROH not arising from recent parental relatedness, particularly on an unpruned basis, whereas after pruning for linkage disequilibrium, such ROH will be discounted. Conversely for African populations long ROH can be expected to arise from more recent parental relatedness, as haplotypes arising from more distant parental relatedness will have been broken down by recombination. As would be predicted we find more unpruned mean SROH for non-African populations, but difference is much less marked for the standard deviation of SROH (Figure 14).

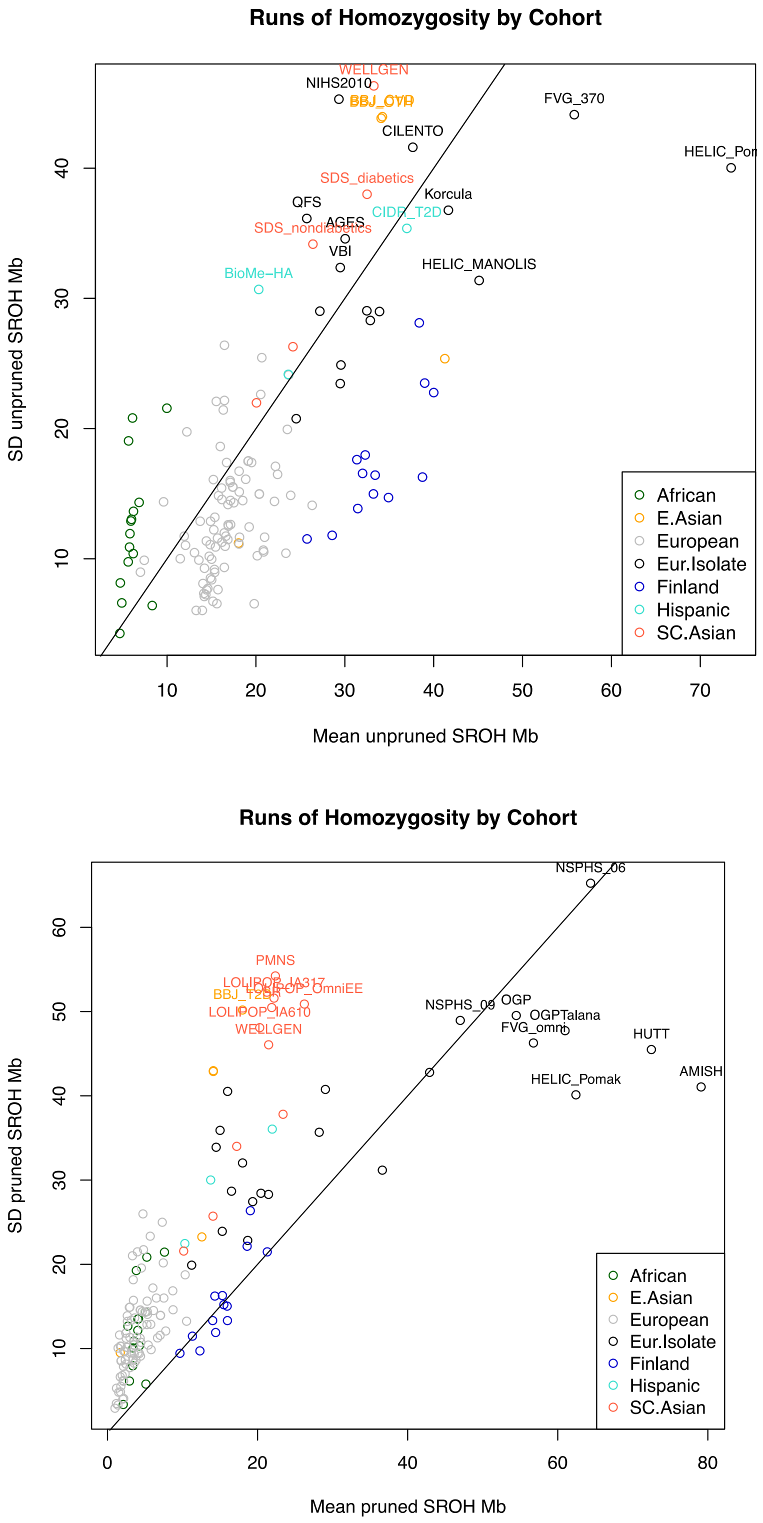
We found that the mean and standard deviation of SROH for a cohort (pruned or unpruned) is related, with a higher coefficient of variation for pruned SROH (points lie further above y=x in Figure 14). As statistical power to detect an association with SROH is driven by variation in SROH, cohort mean homozygosity is thus a broad guide to power (for a given sample size). It is interesting to note that on an unpruned basis, European and African populations separate due to mean SROH, but not the standard deviation of SROH. However the mean separation is much less apparent on a pruned basis (Figure 14). It is also interesting to note the higher coefficient of variation of pruned SROH of Asian relative to other populations.

Figure 13 Runs of Homozygosity by Cohort - mean pruned and unpruned homozygosity.



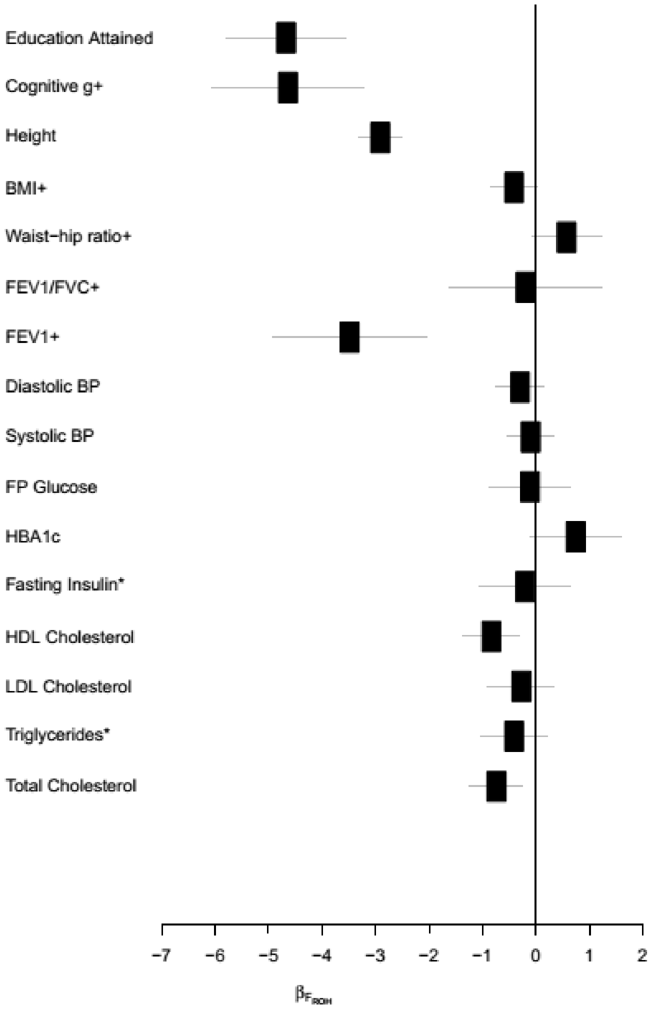
SC.Asian is South & Central Asian, E.Asian is East Asian, Eur.Isolate is European isolates. The ten most homozygous cohorts are labelled: AMISH are the Old Order Amish from Lancaster County, Pennsylvania; HUTT, S-Leut Hutterites from South Dakota; NSPHS, North Swedish Population Health Study, 06 and 09 suffixes are different sampling years from different counties in Northern Sweden; OGP, Ogliastra Genetic Park, Sardinia, Italy; Talana is a particular village in the region; FVG, Friuli-Venezia-Giulia Genetic Park, Italy, omni and 370 suffices refer to subsets genotyped with the Illumina OmniX and 370CNV arrays; HELIC, Hellenic Isolates, Greece, from Pomak villages in Thrace and MANOLIS from Mylopotamos villages in Crete. The coefficient of standard error in cohort mean SROH (CSE =mean/SE of the mean estimate) varies for each cohort. For unpruned SROH it is between 1% and 12%, with a median of 6%. For pruned SROH, it is between 1% and 25% with a median of 6%. The CSE is affected by sample size and variation in SROH relative to the mean (mean and SD of SROH) vary together as illustrated in Figure 14 Runs of Homozygosity by Cohort - relationship between mean and standard deviation.

Figure 14 Runs of Homozygosity by Cohort - relationship between mean and standard deviation – unpruned and pruned



We studied β_{FROH}, (defined as the effect of F_{ROH} on the trait under a linear model) on 16 complex traits of biomedical importance

.



Trait units are intra-sex standard deviations.

β_{FROH} is the estimated effect of F_{ROH} on the trait

95% confidence intervals are also plotted.

+ indicates phenotype was rank transformed

** indicates phenotype was log transformed.*

BMI, body mass index; BP, blood pressure; FP fasting plasma; HbA1c, haemoglobin A1c (glycated haemoglobin); FEV1, forced expiratory volume in one second; FVC, forced vital capacity; HDL, high density lipoprotein; LDL, low density lipoprotein

For height, FEV1 (a measure of lung function), educational attainment (EA) and *g* (a measure of general cognitive ability derived from scores on several diverse cognitive tests), we found the effect sizes were greater than two intra-sex standard deviations (SD), with p-values all less than 10^{-5}. Thus the associations could not plausibly be explained by chance alone, as detailed in Table 12.

Table 12 Effects of genome-wide burden of runs of homozygosity on four traits.

Phenotype	**Outliers**	**Height**	**FEV1+**	**Educational Attainment**	**Cognitive *g+***
Subjects		354,224	64,446	84,725	53,300
P-association	Included	<1 x 10^{-300}	2.1 x 10^{-6}	1.8 x 10^{-10}	2.5 x 10^{-10}
P-heterogeneity	Included	0.014	0.10	1.2 x 10^{-5}	0.071
β_{FROH}-SD	Excluded	-2.91	-3.48	-4.69	-4.64
SE β_{FROH}-SD	Excluded	0.21	0.73	0.58	0.73
β_{FROH}-units	Excluded	-0.188	-2.2	-12.9	-4.64
SE β_{FROH}-units	Excluded	0.014	0.46	1.83	0.73
Units		m	litres	years	SD
First cousin inbreeding depression	Excluded	-1.2	-137	-9.7	-0.29
Units		cm	ml	months	SD

P-association is P value for association, P-heterogeneity is P value for heterogeneity in a meta-analysis between trait and unpruned F_{ROH}, β_{FROH}-SD is the effect size of F_{ROH} on trait expressed in units of intra-sex phenotypic standard deviations and SE is the standard error. β_{FROH}-units is the effect size estimate in the measurement units and SE the standard error. The P values for those traits showing evidence for association are calculated including 5 outlying cohort-specific effect size estimates (an outlier was defined as T-test statistic over 3 for the null hypothesis that the cohort effect size estimate equals the meta-analysis effect size estimate), which is conservative as the majority of these are in the opposite direction. Beta estimates however exclude these outliers, for which there is evidence of discrepancy, and should thus be more accurate. + indicates phenotype was rank transformed; FEV1 is forced expiratory lung volume in one second; g is the general cognitive factor (first unrotated principal component of test scores across diverse domains of cognition).

To ensure that the results were not driven by a few outliers, we repeated the analysis excluding extreme sub-cohort trait results. In all cases, when excluding outliers, the effect sizes and their significance remained similar or increased. After exclusion of outliers, these effect sizes translate into a reduction of 1.2 cm in height and 137 ml in FEV1 for the offspring of first cousins, and into a decrease of 0.3 SD in *g* and 10 months less educational attainment.

Whilst we consider a fixed effect basis to be the most plausible basis for the effect of F_{ROH} we considered the possibility that a random effect was a better model. P-values for heterogeneity of effect sizes were above 1% for all the traits except educational attainment, which had a p-value of 1.2 x $10^{-}5$. Nonetheless, we also analysed the effect of F_{ROH} as a random effect across the meta-analyses (i.e. the effect of F_{ROH} differed in each cohort). Whilst p-values decreased (due to reduced power), they still remained lower than 0.00026. On the other hand, mean effect sizes remained similar or increased, detailed results are shown in Supplement 4.6. So under fixed or random effects meta-analysis, the results are similar and lead to the same conclusion.

To ascertain if a torso size endophenotype underpins both the height and FEV1 signals, and the degree to which the educational attainment and *g* signals are shared, we explored heritabilities and correlations in the ORCADES cohort.

Phenotypic correlations within the cognition- and stature-related traits were 0.37 and 0.43, respectively, whilst phenotypic correlations between the cognitive and stature traits were in the range 0.087-0.176, with genetic correlations exceeding phenotypic correlations and environmental correlations being lower, as shown in Table 13.

Table 13 Heritabilities and phenotypic, genetic and environmental pairwise correlations

		Heritability		Correlation			
Trait 1	Trait 2	**Trait 1**	**Trait 2**	**Phenotypic**	**Genetic**	**Environmental**	**SE Genetic**
height	FEV1	75.9%	43.2%	43.4%	59.7%	24.9%	5.7%
g	EA	54.6%	48.2%	37.0%	59.6%	13.2%	7.4%
height	g	75.9%	56.7%	17.6%	23.0%	8.0%	6.9%
height	EA	76.0%	50.5%	10.1%	14.0%	4.2%	6.9%
FEV1	g	45.1%	56.5%	16.2%	19.1%	13.3%	9.1%
FEV1	EA	45.2%	51.0%	8.7%	27.2%	-8.3%	9.0%

The strong (additive) genetic correlations within stature and cognition traits (both 60%) suggest β_{FROH} is acting on shared endophenotypes within these trait areas. The fact that the FEV1/FVC (forced vital capacity) ratio is not associated with ROH further points to the effect being on lung/chest size rather than airway calibre. The cognition effects cannot be wholly generated by height as an intermediate cause, given the greater effect size for F_{ROH} for cognition. This is emphasised further by lower heritabilities of cognition and the observed low genetic correlations.

Although it has been suggested that β_{FROH*} exceeds β_{FROH} [21], as haplotypes only recently brought into the homozygous state might harbour more deleterious variants, we found no evidence for this; results were similar with either measure, as shown in Table 14.

Table 14 Estimated Beta$_{FROH}$ under pruning and no pruning

F_{ROH}	Edu	Height	FEV1	g
pruned	-5.24	-3.03	-3.67	-5.00
unpruned	-4.69	-2.91	-3.48	-4.64

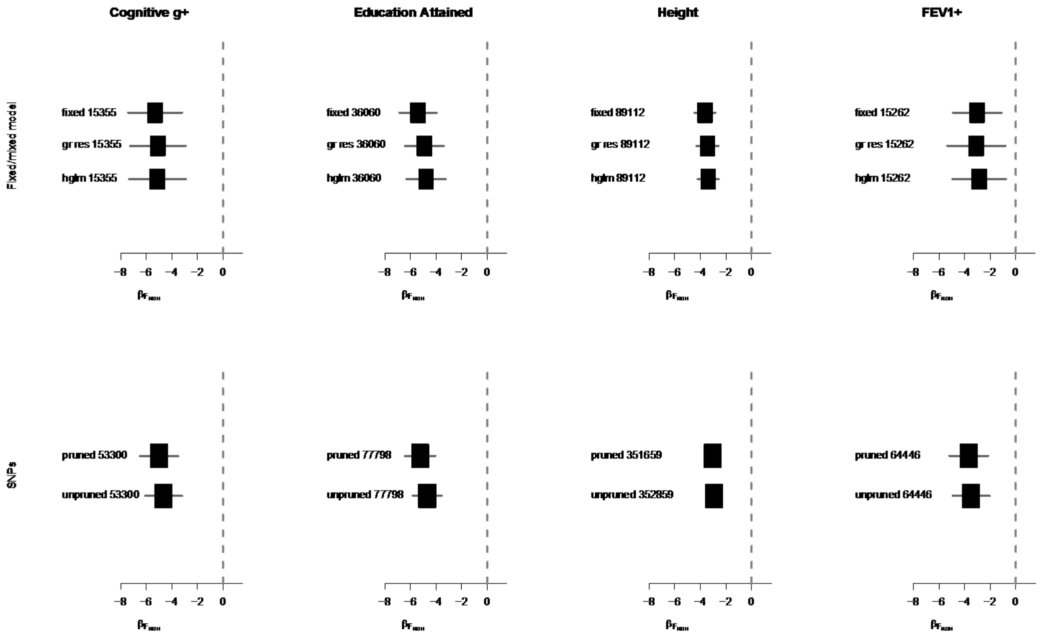
Units are intra-sex trait standard deviation units

We then performed a number of analyses to exclude confounding. As SROH is not heritable (in the narrow sense), a genetic association with population structure or non-sibling relatedness and any heritable trait is not expected as a matter of course. As noted already, we found only small differences (4-12%) between β_{FROH} and β_{FROH*}. This implies that signals of similar strength originate from both ancestral and recent haplotypes, and lends indirect evidence that the observed effects are not due to recent socioeconomic confounding, which would associate with β_{FROH} and β_{FROH*} in different ways. We also found very small differences (3-11% reductions) in estimated β_{FROH}, when comparing the fitting of polygenic mixed models as opposed to fixed-effect-only models, suggesting that polygenic confounding was not substantially affecting the results, as shown in Table 15 and Figure 15.

Table 15 Estimated β_{FROH} under No and different mixed modelling of polygenic (additive) genetic variance

Polygenic effect	Edu	Height	FEV1	g
None	-5.41	-3.63	-3.02	-5.32
Grammar+	-4.91	-3.44	-3.08	-5.08
HGLM	-4.78	-3.37	-2.86	-5.13

Figure 15 Signals of inbreeding depression (β_{FROH})are similar under pruning and fitting polygenic effects as a covariate



gr_res: grammar+ [71] residuals were calculated allowing for (mixed model) polygenic effect of relatedness and other fixed covariates. The effect of F_{ROH} on these residuals was then calculated.

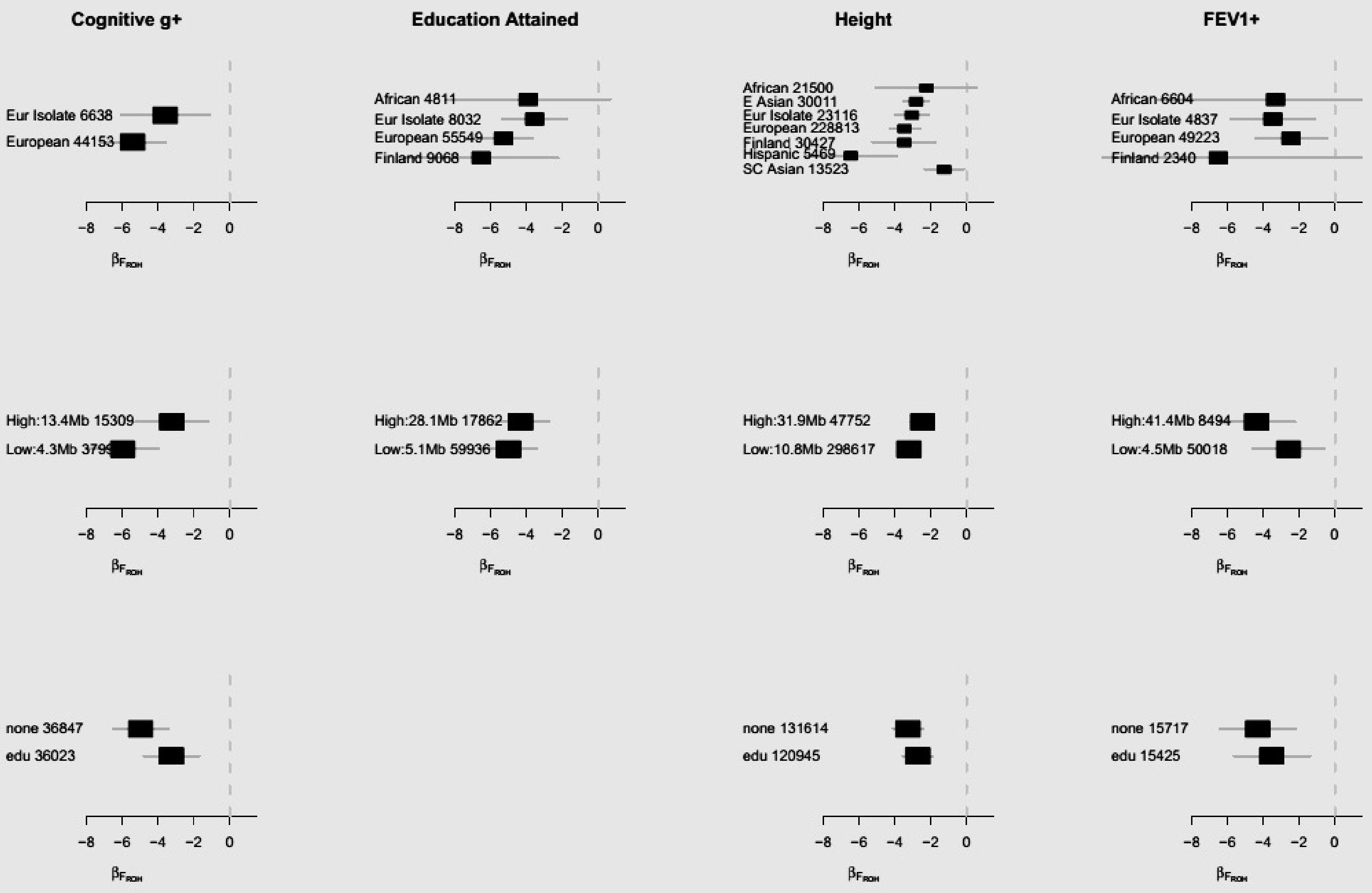
hglm: full hierarchical generalised linear model [76] used to assess effect of of F_{ROH} on trait while simultaneously fitting the (mixed model) polygenic effect of relatedness and other fixed covariates

pruned: SNPs pruned for LD prior to ROH calling

unpruned: Full SNP panel used

The same cohorts are analysed in each row to ensure the effect of the aspect being analysed is the only variable

***Figure 16 Signals of inbreeding depression (β_{FROH}) are robust to stratification by geography or demographic history or inclusion of educational attainment as covariate*.**



Cohorts are divided by continental biogeographic ancestry (African, East Asian, South & Central Asian, Hispanic), with Europeans being divided into Finns, other European isolates (self-declared), and (non-isolated) Europeans. Meta-analysis was carried out for all subsets with 2000 or more samples available, total sample size per subset is given in the plot. β_{FROH} *is consistent across geography and in both isolates and more cosmopolitan populations.*

Cohorts were divided into High and Low ROH strata of equal power and meta-analysis repeated – the effects are consistent across strata for all four traits. The mean SROH is also shown for each stratum in megabases.

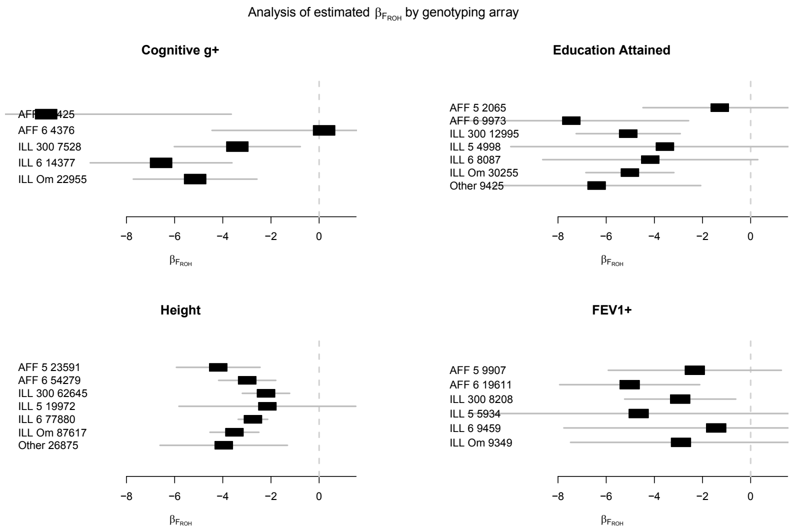
To assess the potential for socio-economic confounding, where available, educational attainment was included in the regression model and compared to a model without educational attainment in the same subset of cohorts. The signals reduce slightly when the education covariate is included; the analysis is not possible for educational attainment as a trait. The numbers indicate the total number of samples in each analysis; they differ because of missing individual educational data within cohorts. + indicates phenotype was rank transformed. FEV1, forced expiratory lung volume in one second; g is the general cognitive component (first unrotated principal component of test scores across diverse tests of cognition); SC Asian is South & Central Asian, E Asian is East Asian, trait units are intra-sex standard deviations and the genomic measure is unpruned SROH.

We continued to evaluate the risk of confounding by conducting stratified and covariate analyses. We found effects of similar magnitude and in the same direction for all four traits across isolated and non-isolated European, Finnish, African, Hispanic, East Asian and South and Central Asian populations

We further tested whether the effect sizes were similar when cohorts were split into more and less homozygous groups. The effect sizes were very similar even though the degree of homozygosity (and variation in homozygosity) varied 3-10-fold between the two strata (depending on which cohorts contributed to the trait). Finally, we fitted educational attainment as a proxy for potential confounding by socio-economic status; this covariate was available in sufficient cohorts to maintain power. The estimated effect sizes for height, FEV1 and *g* all reduced (17%, 18% and 35%, but this might have been expected given the known covariance between these three traits and EA, and the association we found between educational attainment and $F_{ROH.}$ (Figure 16).

Finally, we examined whether estimated effect sizes were influenced by the genotyping array used.

Figure 17 No evidence that estimation of B_{FROH} is influenced by choice of genotyping array



Y axis labels show the genotyping array grouping and the number of subjects measured

AFF5: Affymetrix GeneChip 500k / 5.0 series

AFF6: Affymetrix Genome-Wide Human SNP Array 6.0

ILL 300: Illumina HAP 300/370CNV series

ILL 5 : Illumina HumanHap550 series

ILL 6: Illumina 610/660/670 series

ILL Om: llumina Human Omni series

Although power is limited for the phenotypes with smaller sample sizes, there is no statistically significant evidence that genotyping array influences the estimated effect sizes (95% confidence intervals all overlap), confirming the protocol developed in Chapter 3.

4.4 Discussion

In a very large meta-analysis of up to 350,000 subjects and across 102 diverse cohorts, we find statistically highly significant associations between individual genome-wide summed runs of homozygosity (SROH) and four complex traits: height, forced expiratory lung volume in 1 second (FEV1), general cognitive ability (*g*) and educational attainment. Purely to give a sense an intuitive sense of the magnitude of the effect observed, and on the simplistic assumption that the effects can be extrapolated on a linear basis well outside the observed range of homozygosity, the observed effect sizes would be equivalent to the offspring of first cousins being 1.2 cm shorter and having 10 months less education. Stratified analysis showed consistent effects across continents and cohort mean homozygosity.

Despite the modest reductions in estimated effect sizes for F_{ROH} on height, FEV1 and *g*, when fitting educational attainment as a covariate, the persistence of an effect suggests that most of the signals we observe are genetic. The similarity in effect sizes across continents, consistency of effects with and without fitting relatedness and when using F_{ROH} or F_{ROH*} and in particular in populations with very different degrees of homozygosity, all appear inconsistent with confounding due to environmental or additive genetic effects.

It is also interesting to consider the potential influence of assortative mating, which is commonly observed for human stature, cognition and education[77]. Given the polygenic trait architectures, the phenotypic extremes could be more genetically similar to each other and hence the offspring more homozygous. However, at least in its simplest balanced form, the increase in genetic similarity would be equal at both ends of the phenotypic distribution, leading to no linear association between such genetic similarity and the trait; both tall and short people would be more homozygous. Furthermore, humans also mate assortatively on BMI, for which we see no effect. A more complex possibility, a form of reverse-causality, could arise when subjects from one trait extreme (e.g. more educated people) are on average more geographically mobile, and thus have less homozygous offspring, with those offspring in turn inheriting the trait extreme concerned[78]. We do not think that this mechanism can account for our results, since it does not readily explain the constancy of our results under different models, especially the similarity in β_{FROH} for either more or less homozygous populations. Moreover, we observe similar effects in multiple single village cohorts, and the Amish and Hutterites, where there is no geographic structure and/or no sampling of immigrants, hence such confounding by differential migration cannot occur.

Our results are consistent with previous genomic[21] and pedigree[79] studies, which have shown inbreeding effects on stature with similar effect sizes (0.01 increase in F decreases height by 0.037 SD[79] versus 0.029 SD in the present study). Our genomic confirmation of directional dominance for *g* and discovery of inbreeding effects on educational attainment in a wide range of human populations adds to our knowledge of the genetic underpinnings of cognitive differences, which are currently thought to be largely due to additive genetic effects[80]. Our findings go beyond earlier pedigree-based analyses of recent consanguinity to demonstrate that inbreeding depression is not a result of confounding and influences demographically diverse populations across the globe. The estimated effect size is consistent with pedigree data (0.01 increase in F decreases *g* by 0.046 SD in our analysis and 0.029-0.048 in pedigree-based studies) [81,82]. It is germane to note that one extreme of cognitive function, early onset cognitive impairment, is strongly influenced by deleterious recessive loci[83], so we can speculate that an accumulation of recessive variants of weaker effect may influence normal variation in cognitive function. Although increasing migration and panmixia have generated a secular trend in decreasing homozygosity[84], the Flynn effect, wherein succeeding generations perform better on cognitive tests than their predecessors[85], cannot be explained by our findings, because the intergenerational change in cognitive scores is much larger than the differences in homozygosity would predict. Likewise, the inbreeding effect on height cannot explain a significant proportion of the observed inter-generational increases, nor do inbreeding effects potentially comprise a material part of missing (broad sense) heritability, as the observed variation in homozygosity and the meta-analysed effect sizes, suggest very low proportions of variance explained. For example, the estimated proportion of phenotypic variance explained is 0.01% for height in Generation Scotland, a typical European population, and 0.1% for cognition in ORCADES, a typical genetic isolate.

Inbreeding depression is ubiquitous in plants and is seen for numerous fitness-related traits in animals[86], but no effect was observed for the 12 other mainly cardio-metabolic traits in which variation is strongly age-related. This suggests that previous reports in ecological studies or substantively smaller studies using pedigrees or relatively small numbers of genetic markers may have been false positives[58,59]. The lack of directional dominance on these traits does not, however, rule out many dominant variants, provided dominant variants acting in different directions are cancelling out. ROH analyses within specific genomic regions are warranted to map recessive effects even when there is no genome-wide directional dominance. Such recessive effects have been observed for a subset of cardiovascular risk factors [87] and expression traits [88].

Whilst, directional dominance obviously suggests a search for specific loci where the direction of such dominance arises, it far from precludes the existence of loci with opposite dominance, nor does it preclude the existence of, on average neutral, dominant loci for the other traits. Localised determination of association between runs of homozygosity and phenotype could thus offer an alternative to GWAS and regional heritability [89] to the detection of genes affecting complex traits, and illumination of the genetic architecture underpinning them, although power will depend on the nature of local dominance.

The finding that β_{FROH} and β_{FROH*} appear equal for the four traits also raises interesting questions as to the nature of the selective pressures on these traits – in particular whether the pressures were ancient rather than modern. Equally although our study has shown clear evidence for inbreeding depression measured by ROH, it has not examined whether identity-by-state (IBS) as measured using SNPs, would capture similar or different amounts of inbreeding depression. Indeed a multivariate analysis, which would plausibly be well powered in ROHgen, could well disentangle the effects of IBS, F_{ROH} and F_{ROH*}.

4.5 Conclusions

We have demonstrated the existence of directional dominance or inbreeding depression, on four complex traits (stature, lung function, cognitive ability and educational attainment) whilst showing any effect, if it exists, on the other 12 traits is at least an order of magnitude smaller. This suggests past directional selection for some alleles increasing size and cognition, but not for the other twelve traits, although they are associated with late onset disease in modern settings. The inheritance of complex traits thus has trait specific architecture of subtle, perhaps surprising, forms.

4.6 Supplement: Comparison of fixed and random effects meta-analyses for key phenotypes

Phenotype	n	beta	se_beta	p	effect_type
Cognitiveg+	53300	-4.64	0.73	2.50E-10	fixed
Cognitiveg+	53300	-4.59	1.00	4.37E-06	random
EA	77798	-4.69	0.58	4.44E-16	fixed
EA	77798	-4.96	0.80	6.84E-10	random
Height	352859	-2.91	0.21	<1E-300	fixed
Height	352859	-3.12	0.28	<1E-300	random
FEV1+	64446	-3.48	0.73	2.13E-06	fixed
FEV1+	64446	-3.47	0.95	2.59E-04	random

Chapter 5 Using local exome sequences to impute hidden variants and increase power of Genome Wide Association Studies.

5.1 Introduction

Modern genome wide arrays usually only capture 300k-1M genetic variants, whereas known population variation in the human genome exceeds 37.5m sites [23]. As there is no a priori reason to suspect genome wide arrays to have captured causal loci, causal loci are likely to remain hidden to the researcher using only array data. A commonly used technique in Genome-wide association study (GWAS) is imputation, where further genetic variants known to exist in other populations are inferred into the study population. Meta-analyses, in particular, routinely use genotype imputation, principally to ensure a common panel across studies, but also to give dense coverage[49]. Accurate imputation of less common variants (minor allele frequency MAF, 1-10%) may be particularly useful as commercial genotyping arrays often provide poor coverage of such variants, and imputation improves association power most for less frequent causal variants[24]. Although there is still no certainty that causal variants are captured even by a dense imputation, the likelihood is obviously increased as are the prospects of capturing a variant in strong LD with a causal variant. Imputation and, in particular, dense accurate imputation therefore offers the prospect of getting closer to hidden causal variants.

The recently released 1000 Genomes haplotypes [23] are a particularly large and dense reference panel that will be commonly used as an imputation reference panel, particularly in GWAS consortia. At the same time, theoretical studies and empirical studies using other primary reference panels, have shown that imputation accuracy in a study population can be increased by use of an additional reference panel such as whole genome or exome sequence data drawn from a subset of the population under study [24] [90] [91] [92] [93] [94] [95].

It is therefore useful to quantify the likely benefit of adding local reference data to 1000 Genomes data, particularly for less common variants, and especially if the population is genetically distant from the 1000 Genomes populations.

We used data from the CROATIA-Korcula[2] and Orkney Complex Disease studies (ORCADES) [36] [96]. Both studies are family-based, cross-sectional community studies of the genetics of complex traits. The Croatian island of Korčula is in the Adriatic and the ORCADES study is based in the Orkney Isles in Scotland.

Genotypes obtained from the whole exome sequencing of 91/89 CROATIA-Korcula/ORCADES quality controlled samples were used to supplement the 1000 Genomes reference panel. We focused on less common (MAF 1-10%) exonic variants already in 1000 Genomes which, unlike low frequency, and rare (MAF<1%) or private variants, can be meta-analysed in typically sized consortia.

We thus seek to determine if imputation accuracy can be improved by the addition of local sequences to a global reference panel.

5.2 Method

The ORCADES and CROATIA-Korcula studies both had ethical approval for genetic research into the basis of complex traits, approved by the appropriate committees in each country. For ORCADES the committees were the Orkney Local Research Committee and the North of Scotland Research Ethics Committee (approval Orkney: 27/2/04). For CROATIA-Korcula the committees were the Ethics Committee of the Medical School, University of Split (approval id 2181-198-03-04/10-11-0008) and the NHS Lothian (South East Scotland Research Ethics Committees; REC reference 11/AL/0222). All participants provided written informed consent.

Array genotypes were obtained from Illumina Hap370CNV array, at 319,552 SNPs for CROATIA-Korcula subjects and Illumina Omni1 array at 1,140,419 SNPs or the Illumina Human Hap300 array at 293,687 SNPs for ORCADES subjects. For ORCADES a common panel of intersecting Hap300 and Omni1 SNPs was first

[2] In accordance with the cohort's own convention the cohort name has been spelt without an accent.

created. The panel for CROATIA-Korcula was then restricted to these SNPs, to ensure similar panel sizes.

Subjects to be sequenced were selected from the wider study populations that were genotyped on the Illumina Hap (370CNV/300) arrays to minimize relatedness, and thus to maximize representation of study population haplotypes. The selection was carried out by tracking the identity-by-descent sharing structure, as determined by the array genotypes using the program ANCHAP [97]. Whole exome sequences of 99/95 CROATIA-Korcula/ORCADES subjects were generated using the Agilent SureSelect All Exon 50 Mb kit and 234,746/217,015 variants were identified.

Quality control (QC) of genotyping array data, that were subsequently used for imputation, was in accordance with best practice for association studies[98]

The Korčulan/Orcadian (99/95) exome sequenced subjects' array genotype data was quality controlled alongside the other 801/1069 samples available in each population. 892 Korčulan subjects were genotyped using the Illumina Hap370CNV array, at 319,552 SNPs. Orcadian subjects were genotyped on the Illumina Omni1 array at 1,140,419 SNPs or the Illumina HumanHap300 array at 293,687 SNPs. An intersecting panel of 178,477 SNPs was obtained for 1159 Orcadian subjects.

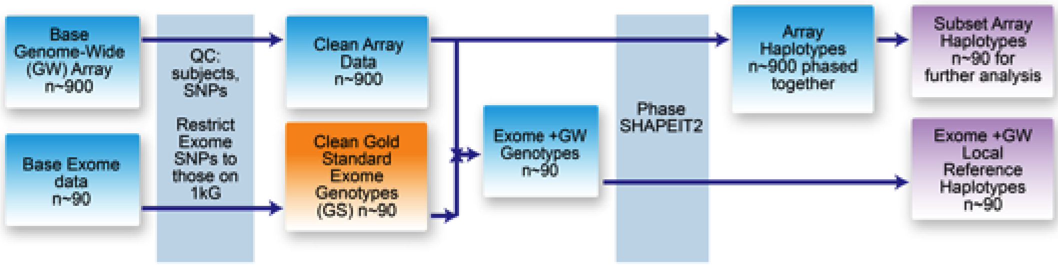
Individuals that failed to genotype at more than 3% of SNPs were excluded, and SNPs that failed to genotype in more than 10% of samples, or failed a test for Hardy-Weinberg equilibrium (p-value=10^{-6}) were excluded. In creating the reference panel, our aim was to create a robust and diverse local panel, so genetic outliers were not excluded, provided that other QC thresholds were satisfactory. Our array data was remapped using LiftOver[99] from NCBI build 36 to build 37.3, to match the exome and 1,000 Genomes data - 9,181/4,607 SNPs were not successfully mapped to the newer build.

1,538/2,115 SNPs and 0/1 samples failed QC, resulting in 892/1158 samples genotyped at 308,833/171,749 SNPs for CROATIA-Korcula/ORCADES. To avoid possible discrepancies associated with a different size of SNP panel, the larger

CROATIA-Korcula panel was then restricted to those intersecting SNPs on the post-QC Orkney panel (but not vice-versa).

As illustrated in Figure 18, post QC array data of 170,134/171,749 SNPs for 892/1158 Korčulan/Orcadian subjects were then pre-phased simultaneously (within each population) using SHAPEIT v1.r416 [100] [101] including the maximal pedigree structure permitted by the software (non-overlapping nuclear families) to create a phased set of study genotypes ready for imputation using IMPUTE2 v2.2.2 [102]. The simultaneous phasing of all (892/1158) study subjects allowed all these subjects' phasing to inform the phase of the ~100 subjects taken forward as a reference panel and for imputation.

Figure 18 Preparation of array data and local reference panel for imputation. The genotype data were quality controlled and phased. These data were then used in further downstream analysis.


Base Genome-Wide (GW) Array n~900
Base Exome data n~90
QC: subjects, SNPs
Restrict Exome SNPs to those on 1kG
Clean Array Data n~900
Clean Gold Standard Exome Genotypes (GS) n~90
Exome +GW Genotypes n~90
Phase SHAPEIT2
Array Haplotypes n~900 phased together
Subset Array Haplotypes n~90 for further analysis
Exome +GW Local Reference Haplotypes n~90

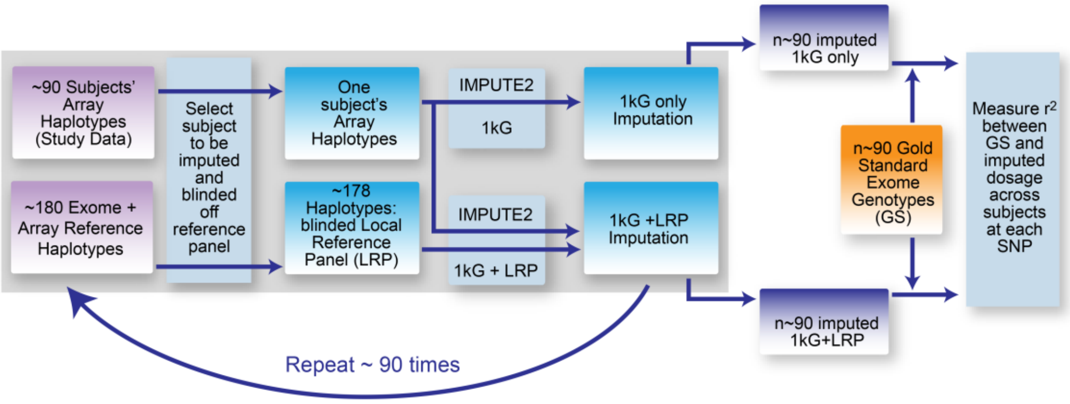
Exome sequence data were also subjected to rigorous QC to ensure they were of high quality so that that the local reference panel we created did not have a significant number of incorrect haplotypes. Variants were called by first aligning the raw sequence data to the human hg19 reference genome using the Stampy short read aligner[103] (with BWA utilized as a pre-mapper[104]). Genotype calls were produced from the resulting alignments using GATK's unified genotyper, following GATK's recommended best practice for variant detection from exome sequence datasets[105]. Variants were required to have a phred-scaled quality of at least 40. Individual sample genotype calls with a phred-scaled quality less than 20 were regarded as missing. Variants that were called in less than 50% of subjects, or with a minor allele frequency of less than 0.75% were removed (hence inclusion required at least two minor alleles across samples). All variants that mapped to more than one homologous region or failed a test of Hardy-Weinberg equilibrium (HWE) with a p-value of less than 10^{-4}, were also removed, leaving 99/95 CROATIA-Korcula/ORCADES subjects genotyped for 102,192/97,052 variants. The HWE test was a more stringent test than for the array data reflecting lower sample numbers and the desire to particularly ensure integrity for reference data. We restricted our analysis to individuals with exome sequences and merged the exomes with the array data for these subjects. Subjects/variants with more than 50/30 mismatching calls, between the array and sequence data were excluded, although no variants failed this test. This resulted in exomes for 93/90 subjects genotyped at 102,192/97,052 exonic SNPs being merged with array data at 170,134/171,749 SNPs for these individuals. The resulting panels had 265,929/262,513 variants which were 99.91%/99.92% concordant, based on the genotypes called on both panels for 6,397/6,285 overlapping variants. As the overall genotypic concordance could mask discrepancies for minor alleles, particularly the less common variants of interest, concordance rates for minor allele calls were calculated in the MAF 1-3% range separately. Only 1/1 (CROATIA-Korcula/ORCADES) call was discrepant on each overlapping panel, giving minor allele concordance of 99.7% in both studies for these variants.

8,150/10,964 Korčulan/Orcadian variants other than single base substitutions, for example insertions or deletions, were excluded. 119/110 conflicting map positions and individuals called at fewer than 80% of the combined SNP panel were then excluded, leaving 91/89 subjects typed across 257,633/251,439 SNPs. Our focus was on the potential to improve power in meta-analyses, so polymorphisms that were unique to each cohort were excluded. This was done by comparison to the 1000 Genomes project map and those variants not present in the 1000 Genomes reference data or with mismatches in allele codes were excluded.

The merged sequence and array data consisting of 233,195/232,096 variants for 91/89 subjects were then phased by SHAPEIT, using the recommended N_e of 11,418 and the default settings [100], to create reference haplotypes, as shown in the lower half of Figure 18.

Having created suitable post-QC array data and secondary reference panels, imputations were performed using genome-wide array data plus (i) 1000 Genomes haplotypes [24] alone or (ii) 1000 Genomes haplotypes together with local data as reference panels. Both imputations were then compared with known genotypes and an assessment of accuracy across all subjects was made for each SNP, as illustrated in Figure 19.

Figure 19 Illustration of the procedure to estimate imputation accuracy.


~90 Subjects' Array Haplotypes (Study Data)
~180 Exome + Array Reference Haplotypes
Select subject to be imputed and blinded off reference panel
One subject's Array Haplotypes
~178 Haplotypes: blinded Local Reference Panel (LRP)
IMPUTE2
1kG
IMPUTE2
1kG + LRP
1kG only Imputation
1kG +LRP Imputation
n~90 imputed 1kG only
n~90 Gold Standard Exome Genotypes (GS)
n~90 imputed 1kG+LRP
Measure r2 between GS and imputed dosage across subjects at each SNP
Repeat ~ 90 times

We used a drop one-out cross-validation approach. For the imputation step each subject was removed from the reference panel in turn, and this subject's exome sequence SNPs were then imputed using either the 1000 Genomes reference panel alone or in conjunction with a second local reference panel. All subjects' imputed allelic dosages were then compared with the exome sequence genotype data ("gold standard").

Imputation of the 91/89 subjects with and without the benefit of local reference data was carried out using IMPUTE2, using the phased reference panel option, the phased array data haplotype option, and with the software splitting the genome into chunks, which had been predetermined to be less than 5Mb in size and avoiding crossing the centromeres. N_e was set to 20,000; all other settings were left at their default values. For the one panel imputation, the 1000 Genomes Phase 1 worldwide integrated variant set (March 2012 release) [23] as available on the IMPUTE2 website [102] was used. The two-panel imputation added the phased local reference data as a secondary panel (we did not use the merge panels option). All other settings for the two-panel imputation were identical to the one panel imputation. We performed imputations for each subject with local exome data separately, with the study subject's own haplotypes removed from the secondary reference panel so that the haplotypes of the individual to be imputed were not present in the reference data. For a given SNP, the accuracy (r^2) of the allelic dosages imputed was measured across samples against the known exome sequence-called genotypes.

As evidenced by the genome-wide SNP array concordance data, noted above, there was close agreement between the exome sequence and independent genotyping data, indicating that the sequences were a suitable gold standard. Furthermore exome array data were also available for the CROATIA-Korcula study (although not ORCADES) and concordance between exome array and exome sequence genotypes was 99.5% and was similar across all MAF bands

The dual use of exome sequences both as a secondary reference panel and as the gold standard to obtain imputation accuracy was considered appropriate since a subject's imputation panel did not include their own sequence, avoiding circularity at the imputation stage.

5.3 Results

We found a significant increase in accuracy (r^2 of imputed against known allele dosages across samples for a given SNP) from use of a local reference panel, which was often substantial for less common variants (Table 16).

Table 16 Mean accuracy of imputation (r^2 of allelic dosage across all samples for a SNP) averaged across SNPs split by Minor Allele Frequency (MAF)

MAF	1-3.2%		3.2-10%		10-32%		>32%	
Population	Korčula	Orkney	Korčula	Orkney	Korčula	Orkney	Korčula	Orkney
N SNPs	12132	12123	11548	10677	16243	15262	10174	9265
r^2 1kG	0.504	0.586	0.729	0.778	0.868	0.894	0.894	0.913
r^2 1kG+LRP	0.697	0.753	0.841	0.867	0.916	0.931	0.934	0.944
Increase r^2	0.193	0.167	0.112	0.089	0.049	0.037	0.039	0.031
Std dev.	0.309	0.295	0.182	0.157	0.093	0.078	0.074	0.065
Inc. Sample	38%	28%	15%	11%	6%	4%	4%	3%

MAF bins increase by factors of √10, to create four exponentially increasing bins.

N SNPs: number of SNPs in MAF bin

1kG: 1000 Genomes used as reference panel

1kG+LRP: 1000 Genomes plus local reference panel

Increase r^2: Average across all SNPs in MAF bin increase in r^2

Std dev: The standard deviation (across SNPs) of the increase in r^2 at each SNP

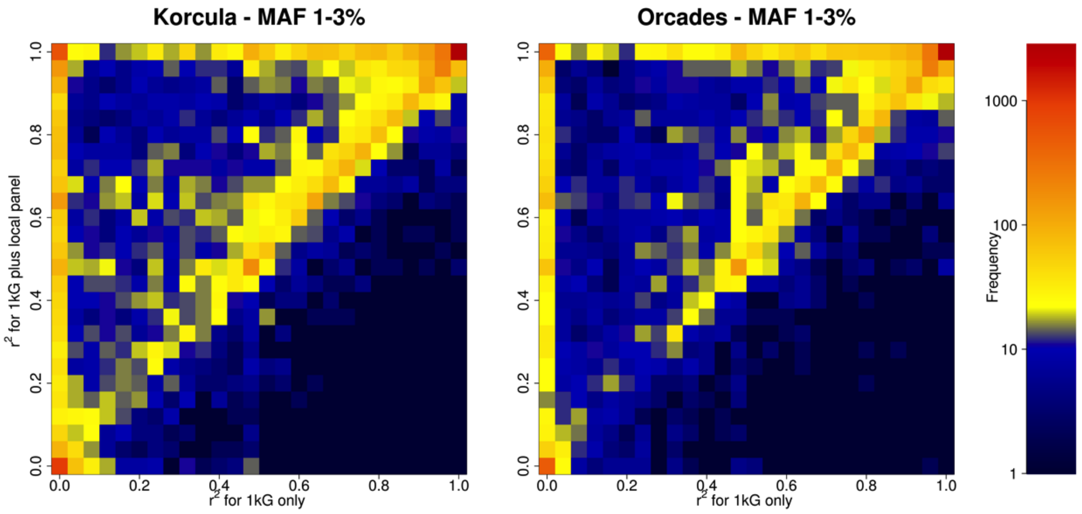
Inc. Sample: Increase in effective sample size for GWAS (=Increase r^2/ r^2 1kG)

The standard errors of mean increases are less than 0.003. All improvements in r^2 are significantly different from zero and significantly different between MAF bands ($P<0.001$, two-sided t tests).

Variants with a minor allele frequency in the range 0.01 – 0.032 showed an increase in imputation accuracy of 0.193/0.167 (38%/28% improvement) for CROATIA-Korcula/ORCADES and 0.112/0.089 (15%/11% improvement) for variants with MAF between 0.032 and 0.100. The high accuracy of the 1000 Genomes imputation for more common variants (MAF >0.1) provided more limited scope for improvement in this category, although even for the most common variants (MAF>0.32) the accuracy of imputation increased by 0.039/0.031 (4%/3% improvement) for CROATIA-Korcula/ORCADES after adding the second (local) reference panel.

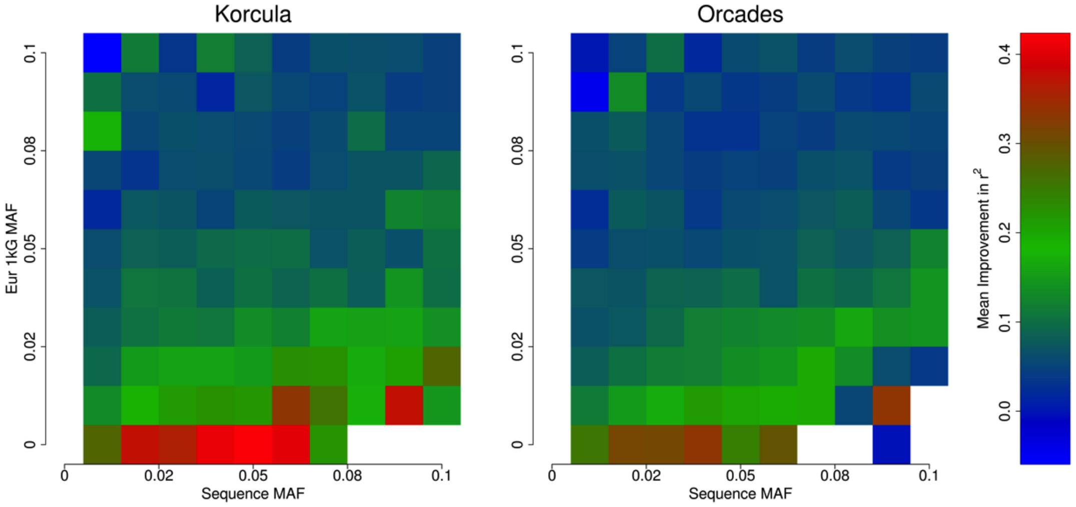
Much of the improvements arise from SNPs that have an r^2 close to zero with the 1000 Genomes-only imputation and which were imputed more accurately with the addition of the local panel (Figure 20). For CROATIA-Korcula/ORCADES 12%/9% of all SNPs imputed poorly (r^2<0.2) using 1000 Genomes data alone. About one-fifth (17.1%/19.9%) of these poorly imputed SNPs imputed well (r^2>0.8) after the addition of the local reference panel.

Figure 20 Frequency plot of imputation accuracy (r^2) using 1000 Genomes data alone against 1000 Genomes plus a local reference panel for SNPs with Minor Allele Frequencies (MAF) of 1-3.2%.



SNPs that were less frequent in 1000 Genomes than in our sequences generally improved more, as illustrated in Figure 21, where areas of greater improvement are generally observed towards the right-hand side in the figure. The effect is more pronounced in Korčula and is particularly marked for variants where MAF is less than 1% on 1000 Genomes European panel.

Figure 21 Plot of mean improvement in imputation accuracy (r^2) for SNPs with minor allele frequency (MAF) in the range 1-10% in our exome sequence data.



Counts of the SNPs in each cell of Figure 21 *are shown in supplementary information at the end of this chapter.*

We also looked at r^2 increase as a function of European 1000 Genomes MAF. As stated above, for SNPs with a MAF of 1-3.2% in our local sequences, the mean increase in r^2 was 0.193/0.167. For these SNPs, the increase in r^2 was 0.297/0.264 for those in the European 1000 Genomes MAF band <1%, 0.137/0.112 for MAF band 1-3.2% and 0.086/0.072 for MAF >3.2%.

5.4 Discussion

Our results show that use of a secondary local reference panel in addition to the 1000 Genomes reference haplotype data can significantly increase the quality of imputations, particularly for less common alleles and the improvement is greater when the study population is genetically further from the populations in the reference data.

We estimated imputation accuracy using a leave-one-out cross-validation approach, in which we compared known genotypes to imputed ones using either the 1000 Genomes reference panel alone or accompanied by a panel obtained from sequence data of individuals from our study populations. Although we took care in our cross-validations to avoid circularity by using the leave-one-out approach in the imputations, for practical reasons, especially computing time, the phasing stage was done only once including all subjects (and therefore included the subject being blinded at the imputation stage). We acknowledge that this could potentially slightly inflate the reported increase in accuracy when using the second reference panel.

Imputation accuracy is not only affected by the quality and composition of the reference data used, but also by the design of the genotyping array, in particular array density and whether the array captures population specific variants [106]. A dense, locally relevant array used to genotype the study population will improve the quality of imputation compared to a less dense one, when using a global reference panel, and thus reduce the potential scope for improvement when adding local sequence data. However, where the study population's haplotypes are distinct, due to recombination, from the reference panel population, the use of a denser array can be expected to improve the imputation but the denser array will also allow even better matching of local haplotypes, and so there should be a further benefit from use of a local secondary reference panel.

Consistent with this hypothesis, the accuracy of base imputations using only the 1000 Genomes reference panel was greater for ORCADES than CROATIA-Korcula, presumably due to the greater proximity of Orkney to subjects in the 1000 Genomes reference panel. Twenty three Orcadians, 77 mainland British and 100 of northern European ancestry individuals are present in the 1000 Genomes data, and principal component analysis shows that Balkan populations (such as Korčula) are more distant from the nearest subjects in 1000 Genomes (Tuscans, from central Italy, N=100), than the variation observed within the British Isles[23] [107]. This suggests to us that, as might be expected, imputation improvement due to addition of local data will be most marked for populations genetically distant from 1000 Genomes samples. Whilst part of the benefit arises from including reference data with allele frequencies closer to the study population, the capture of representative local haplotypes further contributes to the increase in imputation accuracy, and this latter effect will be more marked, or at least require fewer local subject to be sequenced, in isolated populations, where fewer distinct haplotypes will be segregating.

Similarly the much greater improvements in accuracy for SNPs where the MAF is greater in our sequences than 1000 genomes, perhaps not surprisingly, shows that local sequences will add value to imputations in regions of the genome where drift, or other forces, have created a distinct genetic structure.

Comparing these results with those of other researchers who have examined the benefits of study specific reference panels, often using 1000 Genomes like us or HapMap [108] as primary panels, whilst illuminating, is not straightforward. Inevitably, different types and sizes of reference panels are used, as well as different genotyping arrays for the subjects whose genotypes are to be imputed. This is further complicated by different study protocols and differing genetic structure of the study populations. With these caveats, our results of an r^2 of 0.70-0.75 from 90 reference panel subjects in addition to 1000 Genomes seem consistent with those of Liu et al [95] and Auer et al [94], for MAF 1-3%. Neither of these studies used a global reference panel, but Liu et al, in their verification step, attained an r^2 of around 80% with ~2,000 subjects on their (array data) reference panel with unfiltered results, whilst Auer et al obtained an r^2 of 82% with 761 exome reference panel subjects, albeit filtering out lower quality results, using an Rsq threshold of 0.8, where Rsq is equivalent to the squared correlation between nearby imputed and genotyped SNPs [94]. Furthermore the latter study demonstrated that the use of exome imputation can reveal genome-wide significant associations, not discovered by conventional genotyping arrays, as did the study by Holm et al [109], who were able to discern a local rare variant causing sick sinus syndrome, in a large Icelandic study, due to the benefit of adding 87 whole genome sequences to the reference data for their imputation.

Many aspects of our study were similar to a study by Surakka et al [92]. Their Finnish study used 200 (CEU+TSI) HapMap [108] subjects as their primary reference panel and added 81 local subjects genotyped by a genome wide array. For alleles with a MAF <5%, they obtained a median r^2 of 90% for their global panel only imputation rising to 94% after the addition of their local panel. In our study, we report mean r^2, but our median r^2 was 0.77/0.83 rising to 0.88/0.92 after adding the local reference panel for CROATIA-Korcula/ORCADES for a MAF bucket 3-5%. The choice of a 3-5% MAF is intended to correspond to typical array SNPs with MAF<5%. Our results therefore appear consistent with the results of Surakka et al. despite the differences in study design. The study by Uricchio et al [93] obtained much higher mean r^2 (99%), and the technique used for imputation, identifying runs of identity-by-descent (IBD), should be particularly accurate, but its application is restricted to populations which share long haplotypes to a much greater extent than is common in most genetic studies, and we therefore feel our strategy of using 1000 Genomes reference data and adding sequence data from a subset of one's own study subjects is a good practical way forward for many studies.

A proportionate increase in r^2 has the same effect on power as a corresponding increase in study size[110] so the use of high quality sequence data has the potential to provide substantially greater power in GWAS studies for less common variants, particularly those very poorly imputed using 1000 Genomes alone but well imputed with the addition of local exome sequence data.

Our study focused on the exome, but the results should extend to any other genomic region of interest. Moreover, the similar results obtained in our study for two independent populations suggest that corresponding benefits will be found in other studies.

The meta-analysis of multiple populations imputed using local exome sequence data will likely identify new SNP associations. However the amount of variance explained by less common variants individually is likely to be small and will make their detection challenging. This will put increasing emphasis on the use of analytical methods that consider jointly groups of variants, be it gene [111] , regional heritability [89] or network based analyses [112]. Such analyses can also incorporate the potentially valuable information provided by variants private to individual populations including the 24,438/19,343 variants identified by the exome sequencing of the CROATIA-Korcula and ORCADES samples that are not present in 1000 Genomes and hence we have not considered here.

Given the cost and significant practical difficulties in subject recruitment, sequencing a subset of cohort members, for either part or all of the genome, and using these results for imputation will increase power in association studies to discover variants associated with complex traits. Given the density of imputation panels, such imputations will facilitate discovery of variants closer to or perhaps even exactly the elusive causal loci we seek.

5.5 Supplementary Information

5.5.1 Counts of SNPs in each cell underpinning Figure 21

Each cell shows the count of SNPs for the MAF intervals defined by the intersection of row and column headers, where row and column headers represent the lower bounds of those intervals.

CROATIA-Korcula

Local Exome Sequence MAF										
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
1kG MAF										
0.10	2	5	6	19	40	88	114	107	91	130
0.09	4	9	24	26	57	91	111	143	102	107
0.08	3	13	44	49	152	159	139	146	84	106
0.07	10	38	78	86	196	170	151	142	84	90
0.06	25	108	144	128	243	199	169	134	80	72
0.05	79	179	225	180	303	281	168	98	62	33
0.04	205	300	306	202	274	193	117	68	26	23
0.03	560	622	514	288	260	187	67	44	14	12
0.02	955	817	456	178	194	81	28	11	4	2
0.01	2097	1068	523	178	100	39	9	3	3	0
0.00	3933	934	210	52	30	8	0	0	1	0

ORCADES

Local Exome Sequence MAF										
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
1kG MAF										
0.10	1	7	17	22	38	71	91	119	103	121
0.09	3	7	15	56	60	80	111	107	122	158
0.08	3	24	49	66	122	140	164	156	147	153
0.07	8	65	86	101	134	179	178	175	114	90
0.06	45	111	133	144	170	217	189	153	110	68
0.05	105	184	256	203	242	227	217	146	96	55
0.04	195	300	319	246	231	220	125	86	51	28
0.03	544	614	489	380	255	157	92	65	38	23
0.02	1005	740	481	266	170	103	43	31	14	8
0.01	2075	1004	499	251	137	61	22	15	4	3
0.00	3880	1039	296	116	41	22	7	0	0	0

MAF: Minor allele frequency

1kG: 1000 Genomes

Chapter 6 Conclusion

6.1 Findings

My first finding estimated that the genetic basis of human lifespan is less than 16% of the total variation in the trait, somewhat less than previously suggested by the most cited study[19]. I also found that correlations in spouses' lifespans were broadly equal to that between parents and offspring, suggesting that my estimate may be arising from family environment, not genes, and that true heritability may be closer to zero. This has two implications, firstly it re-emphasises the need to control for shared environment when estimating heritability and secondly it suggests that relatively little of the variation in human lifespan can be attributed to genetic causes.

My second finding, was an association, not plausibly explained by chance, between genome-wide homozygosity (as measured by F_{ROH}) and four traits, with an effect size of minus 2-4 phenotypic standard deviations per $F_{ROH.}$ I suggest that genetics is the most plausible explanation for this finding, based on its robustness to stratification by continent and other tests of confounding. I thus showed that cognitive ability and stature were both subject to directional dominance.

Thirdly, I showed that the imputation (estimation) of unmeasured genotypes can be materially improved by using exome sequence data drawn from the local population, over and above that achievable by the currently popular the 1,000 genomes reference panel [23] , and that for conducting a GWAS on less common variants with a minor allele frequency in the range of 1-3%, this improvement in accuracy was equivalent to an increase in the total sample size of 28-38% in the two populations studied.

6.2 Implications

The qualitative result that more data, which are more population-specific, can improve prediction of genotypes, is perhaps, obvious. Nonetheless, the knowledge that a 28-38% increase in effective sample size for GWAS in the MAF range 1-3% can be achieved relatively cheaply by the addition of ~100 local sequences, will guide researchers asking the specific question as to the benefit of such sequence data, when faced with the cost, and often impracticality of recruiting further subjects. More generally the explosion in sequence data in large scale studies, such as UK10K [113] , can not only help trait association analysis in those studies, but also improve imputation, particularly of less common and rare variants, in existing cohorts with genotyping array data, extracting continuing research value for deeply phenotyped studies such as ORCADES [36], with the prospect of on-going discovery or replication of new genetic associations with complex traits.

The lower estimate for the heritability of human longevity than commonly assumed has a number of implications.

Firstly, it suggests that further research is needed to establish robust unbiased estimates of heritability of complex traits in human populations. Robustness will require large sample sizes and powerful study designs (for example thousands of parent-offspring samples). Elimination of bias is more difficult, due to the confounding effects of family environment and socio-economic factors. The studies of Zaitlen at el used extended genealogies [18] and, more recently, admixture mapping [114] in an attempt to reduce confounding of trait and genetic relatedness by within family environment effects and both produced significantly reduced estimates of heritability. Further work is needed, both in terms of replication and to test further methods to control for environmental associations with relatedness, such as careful measurement of confounding factors as environmental correlations with relatedness may still be present in more distant relationships. If true heritability for many complex traits is generally lower than commonly reported, this has simple and obvious consequences for the question of missing heritability [12] – not as much is missing as was thought. But also for the search for causal variants – present commonly used genotyping arrays are capturing an even greater proportion of the (additive) genetic variation than presently believed [17]. And thus larger and larger GWAS using existing arrays can be expected to explain a larger proportion of the true heritability [17], whilst the search for variants not well tagged by present genotyping arrays may be less fruitful, or at least explain less variance than is presently supposed.

Secondly, the limited genetic role suggests that, improvements in human lifespan within the current observed range can be achieved through environmental interventions, although chance will inevitably continue to play a role.

Thirdly, it suggests that conventional GWAS is going to find the search for longevity variants a particularly hard task. In any case, the search for causal variants is exacerbated by low sample sizes used in longevity studies (a few thousand) [115] compared with, for example, height (more than 700,000) [116] due to the difficulty of collection and the fact that in typical population-based cohorts, such as ORCADES[36], most subjects were recruited in middle age and are still very much alive. All of which perhaps explains why only two variants associated with longevity have thus been reliably replicated [117] and a recent meta-analysis with 20,000 long-lived cases found only one new variant (on chromosome 5) associated with long-livedness [118].

Fourthly, and perhaps most interestingly, researchers estimate susceptibility to individual killer diseases [44] [45] [46] is higher than our estimate for lifespan, although one recent study did suggest that a polygenic score based on disease susceptibility did weakly associate with short term survival [119]. This calls into question a naïve assumption that variation in longevity mainly arises due to variation in killer disease susceptibilities and leaves much room for biomedical factors affecting lifespan beyond those of immediate clinical relevance.

Finally, individualised prediction of human longevity may better focus on environmental factors, or biomarkers of ageing [48] rather than genetic factors set at birth, confirming the case for lifestyle and medical interventions to improve long-livedness.

The perhaps surprising lack of heritability of human lifespan, initially suggests that the search for the genetic basis of mortality is a pointless task. However, limited heritability itself raises a number of interesting questions. Are there complicated genetic effects on death and disease, which have limited or no overall effect? For example are genetic risk factors for one disease protective against other diseases due to antagonistic pleiotropy [120], or more speculatively still are risk factors for mortality at one stage in life, protective at other stages? This would raise interesting questions about the causes of disease and the efficacy of its prevention. Both such hypotheses could account for a genetic basis for disease susceptibility, but not overall lifespan, as could simpler models such as environmentally driven frailty preceding partly genetically caused subsequent disease, or the statistical nature of combination of somewhat dependent individual environmental and somewhat independent genetic risks. Research into these questions could proceed along the following lines. Studies of all cause mortality using cox models, could be made using different hazard ratios for different decades of age. Similarly all cause mortality analysis could be tested for association with known genetic risk factors for disease on SNP at a time, or using polygenic risk scores [121] for specific diseases, or all major causes of death simultaneously. In any case, my result suggesting limited heritability of lifespan, even if verified, still clearly leaves open much space for understanding the genetic and wider bio-medical basis of ageing and lifespan, two of the most interesting and complex traits.

The presence of inbreeding depression for size and cognition is proposed to arise either due to deleterious partly recessive alleles or overdominance and recent work suggests the deleterious partly recessive hypothesis is more supported by the evidence [20]. Thus decreases in these traits have been disfavoured by evolution, with purifying selective pressures keeping recessive alleles at low frequency but not eliminating them. This suggests the existence of (recessive) rare alleles perhaps of moderate effect which may form a material part of the genetic basis of these traits and more generally, the existence of dominance variation for these traits, suggests that it may exist for other traits, even those without directional dominance, although even for height GWAS has yet to reveal strong evidence of recessive loci [116]. The genetic covariance between body size and cognition also suggests an underlying trait of generalised healthy development, linked to evolutionary fitness, again for which there may be directional dominance, and plausible association with evolutionary fitness. If true, focus on such variants could prioritise research into pathways of particular interest to developmental biologists.

Multiple avenues for further research using ROH into are suggested. Firstly, the extension to additional traits, especially fitness-related ones such as fertility is an obvious priority. Methodological extensions are also called for, for example looking at different lengths of ROH, pruned or unpruned ROH, and perhaps also just considering heterozygosity as measured directly by an array – which will measure IBD over a short genomic segment. Indeed a multivariate analysis of this form, could well reveal precisely what sort of homozygosity is leading to the association, and facilitate exploration of the other side of the coin – i.e. the benefits of increased heterozygosity. A further methodological extension could consider whether LD pruning is the best way to distinguish common and rare haplotypes. At the same time, and perhaps of greater bio-medical relevance, β_{FROH} could be measured on a regional, rather than genome-wide basis, giving indications of genes, or other functional elements that are associated with dominance. Power considerations will be important here: on the one hand F_{ROH} measured within a small region ($F_{ROH:region}$)is likely to vary fully across [0:1], indeed it may well be a binary trait depending on the length of the region, but on the other hand for a specific region most subjects will have $F_{ROH:region} = 0$. Study of $F_{ROH:region}$ will thus have power aspects in common with those of rare variants, and again as with rare variants, researchers would have the hope that large effect sizes, might be detectable, even after compensation for multiple testing. Given the power considerations, meta-analysis will be particularly important. A first analysis could study $\beta_{FROH:region}$. Development of a uniform protocol, as adopted in ROHgen, will be important, most obviously in terms of specifying precisely each region to be considered and developing a reporting protocol to enable centralised meta-analysis, in much the same way as GWAMA studies do for imputed SNPS. However, further research into the optimal region size and perhaps nature (eg genic) is needed first, balancing issues of multiple testing, haplotype length, and allelic heterogeneity between studies. Optimisation of ROH calling parameters for regions, rather than the whole genome should also be considered. Having found such regions, further analysis could then reveal whether particular haplotypes carry the causal (recessive) allele and fine mapping might reveal it. Whilst all populations should be amenable to ROH-trait analyses, populations in which ROH are highly variable (in particular South and West Asian ancestries, Latino/Hispanic Americans, isolated populations and populations in which consanguinity is practiced) will be most powerful. Analysis of ROH thus offers the prospect of theoretical and biological insights into the nature of complex traits.

The existence of directional dominance and the inference of rare non-additive alleles of moderate effect suggests that revisiting estimation of dominance variation in human populations, using modern techniques might be of value. The confounding effect of common family environment is well known as a source of bias in dominance variation estimates, [6], but large studies such as Generation Scotland [122] and UK Biobank may enable more careful control of such bias, particularly by the application of modern genomic techniques using unrelated individuals [17], which would avoid the effect of family environment, at the expense of a much less powerful study design, requiring the large sample sizes that are only just becoming available. An alternative approach could investigate to what extent h^2_{GWAS} (the genetic variance explained by robustly identified genetic variants) can be increased by inclusion of a dominance term in the genetic model, most obviously to the ~700 known variants affecting human height [116], although this would require robust establishment of the existence of dominance first and then unbiased estimates of the dominance components at each GWAS SNP. Our result finding directional dominance using ROH thus appears to be the end of the (admittedly long) beginning of homozygosity studies in humans, rather than the beginning of the end.

6.3 Final Words

The unprecedented scope and scale of data becoming available to researchers through whole genome sequencing, and projects such as UK Biobank, means that the benefits to ordinary people from the genomics era will eventually fulfil its potential, while at the same time giving the sheer joy of many new discoveries to researchers. In the words of Cynthia Morton, 2014 President of the American Society of Human Genetics and my full agreement, human geneticists are "having the time of our lives" [123] .

Bibliography

1. Mayeux R (2005) Mapping the new frontier: complex genetic disorders. J Clin Invest 115: 1404-1407.

2. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-838.

3. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

4. Galton F (1886) Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland Vol. 15: 246-263.

5. Fisher RA (1918) The correlation between relatives

under the supposition of Mendelian inheritance. Transactions

of the Royal Society of Edinburgh 52: 399-433.

6. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Burnt Mill, Harlow, Essex, England

New York: Longman

Wiley. xii, 438 p. p.

7. Hartl DC, A (2007) Principles of Population Genetics. London: Sinaeur.

8. NCBI (2013) Online Mendelian Inheritance in Man.

9. Human Genome Program (2008) Genomics and Its Impact on Science and Society. In: Energy USDo, editor.

10. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

11. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478: 103-109.

12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

13. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7: e1002051.

14. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177-182.

15. Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2: e41.

16. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4: e1000008.

17. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.
18. Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, et al. (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet 9: e1003520.
19. Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B, et al. (1996) The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. Hum Genet 97: 319-323.
20. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet 10: 783-796.
21. McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, et al. (2012) Evidence of inbreeding depression on human height. PLoS Genet 8: e1002655.
22. Bateson W, et al. . Experimental studies in the physiology of heredity.; 1905.
23. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
24. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906-913.
25. Beeton M, Pearson K (1899) Data for the problem of evoluation in man. II A first study on the inheritance of longevity and the selective death rate in man. Proceedings Royal Society B 67: 290-305.
26. Cohen BH (1964) Family Patterns of Mortality and Life Span. Q Rev Biol 39: 130-181.
27. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
28. Tenesa A, Haley CS (2013) The heritability of human disease: estimation, uses and abuses. Nat Rev Genet 14: 139-149.
29. McGue M, Vaupel JW, Holm N, Harvald B (1993) Longevity is moderately heritable in a sample of Danish twins born 1870-1880. J Gerontol 48: B237-244.
30. Mitchell BD, Hsueh WC, King TM, Pollin TI, Sorkin J, et al. (2001) Heritability of life span in the Old Order Amish. Am J Med Genet 102: 346-352.
31. Ljungquist B, Berg S, Lanke J, McClearn GE, Pedersen NL (1998) The effect of genetic factors for longevity: a comparison of identical and fraternal twins in the Swedish Twin Registry. J Gerontol A Biol Sci Med Sci 53: M441-446.
32. Iachine IA, Holm NV, Harris JR, Begun AZ, Iachina MK, et al. (1998) How heritable is individual susceptibility to death? The results of an analysis of survival data on Danish, Swedish and Finnish twins. Twin Res 1: 196-205.
33. Schoenmaker M, de Craen AJ, de Meijer PH, Beekman M, Blauw GJ, et al. (2006) Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. Eur J Hum Genet 14: 79-84.
34. Gudmundsson H, Gudbjartsson DF, Frigge M, Gulcher JR, Stefansson K (2000) Inheritance of human longevity in Iceland. Eur J Hum Genet 8: 743-749.

35. Montesanto A, Latorre V, Giordano M, Martino C, Domma F, et al. (2011) The genetic component of human longevity: analysis of the survival advantage of parents and siblings of Italian nonagenarians. Eur J Hum Genet 19: 882-886.
36. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, et al. (2008) Runs of homozygosity in European populations. Am J Hum Genet 83: 359-372.
37. R Core Team (2012) R: A language and environment for statistical

computing. . Vienna, Austria.: R Foundation for Statistical Computing,.
38. Lumley T (2012) rmeta: Meta-analysis.
39. Beekman M, Blanche H, Perola M, Hervonen A, Bezrukov V, et al. (2013) Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. Aging Cell 12: 184-193.
40. Gavrilov LA, Gavrilova NS (2012) Biodemography of exceptional longevity: early-life and mid-life predictors of human longevity. Biodemography Soc Biol 58: 14-39.
41. Cournil A, Legay JM, Schachter F (2000) Evidence of sex-linked effects on the inheritance of human longevity: a population-based study in the Valserine valley (French Jura), 18-20th centuries. Proc Biol Sci 267: 1021-1025.
42. Crawford MH, Rogers L (1982) Population genetic models in the study of aging and longevity in a Mennonite community. Soc Sci Med 16: 149-153.
43. Philippe P, Opitz JM (1978) Familial correlations of longevity: An isolate-based study. Am J Med Genet 2: 121-129.
44. Fischer M, Broeckel U, Holmer S, Baessler A, Hengstenberg C, et al. (2005) Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. Circulation 111: 855-862.
45. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, et al. (2006) Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 63: 168-174.
46. Czene K, Lichtenstein P, Hemminki K (2002) Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer 99: 260-266.
47. de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9: e1003608.
48. Fischer K, Kettunen J, Wurtz P, Haller T, Havulinna AS, et al. (2014) Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. PLoS Med 11: e1001606.
49. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, et al. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17: R122-128.
50. Keller MC, Simonson MA, Ripke S, Neale BM, Gejman PV, et al. (2012) Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS Genet 8: e1002656.
51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

52. Howrigan DP, Simonson MA, Keller MC (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genomics 12: 460.
53. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
54. International HapMap C (2003) The International HapMap Project. Nature 426: 789-796.
55. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, et al. (2010) Genomic runs of homozygosity record population history and consanguinity. PLoS One 5: e13996.
56. Garrod AE (1902) About Alkaptonuria. Med Chir Trans 85: 69-78.
57. Dawrwin C (1876) The effects of cross and self-fertilisation in the vegetable kingdom. London: John Murray.
58. Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, et al. (2003) Inbreeding and the genetic complexity of human hypertension. Genetics 163: 1011-1021.
59. Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, et al. (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet 16: 233-241.
60. Wright S (1977) Wright, S. Evolution and the Genetics of Populations, Vol. 3: Experimental Results and Evolutionary Deductions: University of Chicago Press.
61. S W (1922) Coefficients of inbreeding and relationships. Am Nat 56: 330–339.
62. Bittles AH, Neel JV (1994) The costs of human inbreeding and their implications for variations at the DNA level. Nat Genet 8: 117-121.
63. Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65: 1493-1500.
64. Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15: 789-795.
65. Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics 189: 237-249.
66. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, et al. (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 44: 991-1005.
67. Global Lipids Genetics C, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, et al. (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45: 1274-1283.
68. Obeidat M, Wain LV, Shrine N, Kalsheker N, Soler Artigas M, et al. (2011) A comprehensive evaluation of potential lung function associated genes in the SpiroMeta general population sample. PLoS One 6: e19382.
69. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, et al. (2013) GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340: 1467-1471.

70. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294-1296.
71. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components-based method for whole-genome association analysis. Nat Genet 44: 1166-1170.
72. Ronnegard L, Shen X, Alam M (2010) hglm: A Package for Fitting Hierarchical Generalized Linear Models. R Journal 2: 20-28.
73. Marioni RE, Davies G, Hayward C, Liewald D, Kerr SM, et al. (2014) Molecular genetic contributions to socioeconomic status and intelligence. Intelligence 44: 26-32.
74. Hedges LO, I. (1985) Statistical Methods for Meta-Analysis New York: Academic Press
75. Mellars P (2006) Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proc Natl Acad Sci U S A 103: 9381-9386.
76. Shen X, Ronnegard L, Carlborg O (2011) Hierarchical likelihood opens a new way of estimating genetic values using genome-wide dense marker maps. BMC Proc 5 Suppl 3: S14.
77. Mascie-Taylor CG, Boldsen JL (1988) Assortative mating, differential fertility and abnormal pregnancy outcome. Ann Hum Biol 15: 223-228.
78. Abdellaoui A (in submission) Educational Attainment Influences Genetic Variation through Migration and Assortative Mating. PLoS One.
79. Schull WJ (1962) Inbreeding and maternal effects in the Japanese. . Eugen Quart 9.
80. Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, et al. (2012) Genetic contributions to stability and change in intelligence from childhood to old age. Nature 482: 212-215.
81. Morton NE (1978) Effect of inbreeding on IQ and mental retardation. Proc Natl Acad Sci U S A 75: 3906-3908.
82. Bashi J (1977) Effects of inbreeding on cognitive performance. Nature 266: 440-442.
83. Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, et al. (2011) Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478: 57-63.
84. Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, et al. (2009) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 5: e1000415.
85. Flynn JR (1987) Massive IQ gains in 14 nations: what IQ tests really measure. Psychol Bull 101: 171–191
86. Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, et al. (2014) High-throughput sequencing reveals inbreeding depression in a natural population. Proc Natl Acad Sci U S A 111: 3775-3780.
87. Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H (2003) A polygenic basis for late-onset disease. Trends Genet 19: 97-106.
88. Powell JE, Henders AK, McRae AF, Kim J, Hemani G, et al. (2013) Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. PLoS Genet 9: e1003502.

89. Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2012) Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. PLoS One 7: e46501.

90. Zeggini E (2011) Next-generation association studies for complex traits. Nat Genet 43: 287-288.

91. Jewett EM, Zawistowski M, Rosenberg NA, Zollner S (2012) A coalescent model for genotype imputation. Genetics 191: 1239-1255.

92. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, et al. (2010) Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res 20: 1344-1351.

93. Uricchio LH, Chong JX, Ross KD, Ober C, Nicolae DL (2012) Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. Genet Epidemiol 36: 312-319.

94. Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, et al. (2012) Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am J Hum Genet 91: 794-808.

95. Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, et al. (2012) Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. Genet Epidemiol 36: 107-117.

96. Polasek O, Marusic A, Rotim K, Hayward C, Vitart V, et al. (2009) Genome-wide association study of anthropometric traits in Korcula Island, Croatia. Croat Med J 50: 7-16.

97. Glodzik D, Navarro P, Vitart V, Hayward C, McQuillan R, et al. (2013) Inference of identity by descent in population isolates and optimal sequencing studies. Eur J Hum Genet.

98. Weale ME (2010) Quality control for genome-wide association studies. Methods Mol Biol 628: 341-372.

99. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34: D590-598.

100. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9: 179-181.

101. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955-959.

102. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.

103. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936-939.

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

105. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

106. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499-511.

107. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101.

108. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

109. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, et al. (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 43: 316-320.

110. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1-14.

111. Huang H, Chanda P, Alonso A, Bader JS, Arking DE (2011) Gene-based tests of association. PLoS Genet 7: e1002177.

112. Cabrera CP, Navarro P, Huffman JE, Wright AF, Hayward C, et al. (2012) Uncovering networks from genome-wide association studies via circular genomic permutation. G3 (Bethesda) 2: 1067-1075.

113. Muddyman D, Smee C, Griffin H, Kaye J (2013) Implementing a successful data-management framework: the UK10K managed access model. Genome Med 5: 100.

114. Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, et al. (2014) Leveraging population admixture to characterize the heritability of complex traits. Nat Genet 46: 1356-1362.

115. Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, et al. (2011) Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. Aging Cell 10: 686-698.

116. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet advance online publication.

117. Deelen J, Beekman M, Capri M, Franceschi C, Slagboom PE (2013) Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. Bioessays 35: 386-396.

118. Deelen J, Beekman M, Uh HW, Broer L, Ayers KL, et al. (2014) Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. Hum Mol Genet 23: 4420-4432.

119. Ganna A, Rivadeneira F, Hofman A, Uitterlinden AG, Magnusson PK, et al. (2013) Genetic determinants of mortality. Can findings from genome-wide association studies explain variation in human mortality? Hum Genet 132: 553-561.

120. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607-618.

121. Dudbridge F (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genet 9: e1003348.

122. Kerr SM, Campbell A, Murphy L, Hayward C, Jackson C, et al. (2013) Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. BMC Med Genet 14: 38.

123. Morton C. The time of our lives; 2014; San Diego.

Appendix 1a ROHgen Participating cohorts

Cohort	Full name	PMID reference
AEGS1	Athero-express biobank study	15678794
AEGS2	Athero-express biobank study	15678794
AGES-Reykjavik	Age gene/environment susceptibility Reykjavik Study	17351290
AIDHS/SDS	Sikh Diabetes Study	23300278
ALSPAC	Avon Longitudinal Study of Parents and Children	22507742
AMISH	Amish Heredity and Phenotype Intervention Heart Study	18440328, 18805900
ARIC	Atherosclerosis Risk in Communities	2646917
ASCOT-SC	Anglo-Scandinavian Cardiac Outcomes Trial	11685901
ASCOT-UK	Anglo-Scandinavian Cardiac Outcomes Trial	11685901
ASPS	Austrian Stroke Prevention Study	7800110, 10408549
ASPS-Fam	Austrian Stroke Prevention Study	
B58C	British 1958 Birth Cohort	17255346
BASEII	Berlin Aging Study	23505255
BBJ	Biobank Japan	24390342
BRIGHT	British Genetics of Hypertension	12826435
CARDIA_CARe_AA	Coronary Artery Risk Development in Young Adults Candidate Gene Association Resource	3204420
CARL	Carlantino Study	
CHOP - African American	Children's Hospital of Philadephia	
CHOP - Caucasian	Children's Hospital of Philadephia	
CHS (African American)	Cardiovascular Health Study	1669507
CHS (European)	Cardiovascular Health Study	1669507
CIDR_T2D	Starr County Health Studies' Genetics of Diabetes Study	6637993
CILENTO	Cilento Study	17476112 , 19550436
CLHNS	Cebu Longitudinal Health and Nutrition Study	20507864
COLAUS	Cohort Lausannois	18366642

COPSAC2000	Copenhagen Prospective Study on Asthma in Childhood	15521375
Corogene case	Corogene study	21642350
Corogene control	Corogene study	19959603
CROATIA_Korcula	Croatia-Korcula study	24170729
CROATIA_Split _370CNV	Croatia-Split study	24170729
CROATIA_Split+OMNIX+	Croatia-Split study	24170729
CROATIA_Vis	Croatia-Vis Study	24170729
DESIR	Data from an Epidemiological Study on the Insulin Resistance syndrome	8927780
DNBC	Danish National Birth Cohort	11775787
EGCUT_370	Estonian Genome Centre University of Tartu	24518929
EGCUT_OMNI	Estonian Genome Centre University of Tartu	24518929
eMERGE_PAD	Genome-Wide Association Study of Peripheral Arterial Disease	18176561
EPIC	European Prospective Investigation into Cancer-Norfolk	10466767
ERF	Erasmus Rucphen Family Study	15845033
FamHS Human 1M-Duov3	Family Heart Study	11713718
FamHS Human 6100-Quadv1	Family Heart Study	11713718
FamHS HumHap550K	Family Heart Study	11713718
Fenland	Fenland Study	20935629
FHS	Framingham Heart Study	14819398; 1208363; 17372189
FINRISK/ENGAGE	Finrisk study	19959603
FTC_1	Finnish Twin Cohort	23298696,17254406,12537860,12537859
FTC_2	Finnish Twin Cohort	23298696,17254406,12537860,12537859
FTC_3	Finnish Twin Cohort	23298696,17254406,12537860,12537859
FUSION	Finland-United States Investigation of NIDDM Genetics	17463248
FVG	Genetic Park Friuli Venezia Giulia	
GeneSTAR AA	GeneSTAR study	3170971; 2128738; 16551714
GeneSTAR EA	GeneSTAR study	3170971; 2128738; 16551714
GENOA (African American)	Genetic Epidemiology Network of Arteriopathy	15121494
GENOA (European)	Genetic Epidemiology Network of Arteriopathy	15121494

GoDARTS	Genetics of Diabetes Audit and Research in Tayside Scotland	22456734
GOLDN	Genetics of Lipid Lowering Drugs and Diet Network	17446329
GOYA	Genetics of Obesity in Young Adults	21935397
GRAPHIC	Genetic Regulation of Ambulatory Blood Pressure in the Community	2253977
GS:SFHS	Generation Scotland: Scottish Family Health Study	22786799
H2000/Genmets case	Health 2000	
H2000/Genmets control	Health 2000	
HBCS	Helsinki Birth Cohort Study	21613556
HELIC_MANOLIS	Hellenic Isolates - Minoan Isolates	24343240
HELIC_POMAK	Hellenic Isolates - Pomak	24343240
HPFS-Affymetrix	Health Professionals Follow up Study	
HPFS-Illumina	Health Professionals Follow up Study	
HPFS-Omni	Health Professionals Follow up Study	
HRS	Health and Retirement Study	24671021
HRS	Health and Retirement Study	24671021
HUFS	Howard University Family Study	19609347
Hutterites	Hutterites study	23932459
HyperGEN - African Americans	HyperGEN study	10964005
HyperGEN - Caucasians	HyperGEN study	10964005
HYPERGENES - Normotensives	European Network for Genetic-Epidemiological Studies	22184326
InCHIANTI	Invecchiare in Chianti	11129752
INCIPE	Initiative on Nephropathy, of relevance to public health, which is Chronic, possibly in its Initial stages, and carries a Potential risk of major clinical Endpoints	
INCIPE2	Initiative on Nephropathy, of relevance to public health, which is Chronic, possibly in its Initial stages, and carries a Potential risk of major clinical Endpoints	
INDICO_case	Indian Diabetes Consortium	23209189
INDICO_control	Indian Diabetes Consortium	23209189

IPM-BioMe	Institute for Personalized Medicine BioMe biobank	23583978
JHS	Jackson Heart Study	16320381
KORA F3	Cooperative Health Research in the Region of Augsburg	16032513, 16032514
KORA F4	Cooperative Health Research in the Region of Augsburg	16032513, 16032514
LBC1921	Lothian Birth Cohort 1921	14717632 and 22253310
LBC1936	Lothian Birth Cohort 1936	18053258 and 22253310
LHS_EA	Lung Health Study	8500311
LOLIPOP_EW610	London Life Sciences Prospective Population Study	21909115
LOLIPOP_EWA	London Life Sciences Prospective Population Study	18940312
LOLIPOP_EWP	London Life Sciences Prospective Population Study	18193046
LOLIPOP_IA317	London Life Sciences Prospective Population Study	18454146
LOLIPOP_IA610	London Life Sciences Prospective Population Study	19820698, 19651812
LOLIPOP_IAP	London Life Sciences Prospective Population Study	18193046
LOLIPOP_OmniEE	London Life Sciences Prospective Population Study	23222517
MAYWOOD	Maywood Study	8793366; 20400458
MEGA FU	Multiple Environmental and Genetic Assessment	22253578 / 15701913
MESA	Multi-ethnic Study of Atherosclerosis	12397006
METSIM	METabolic Syndrome In Men	19223598
MICROS	Micro-isolates in South Tyrol	17550581
NHS-Affymetrix	Nurses Health Study	
NHS-Illumina	Nurses Health Study	
NHS-Omni	Nurses Health Study	
NIGERIA	Nigerian Study	8880560; 20400458; 22615923
NIHS	Norfolk Island Health Study	24314549
NSPHS_06	North Swedish Population Health Study	
NSPHS_09	North Swedish Population Health Study	

NTR	Netherlands Twin Registry	23186620; 23298648
OBA	French Adult Obese	19151714
OGP	Ogliastra Genetic Park	19247500
OGP Talana	Ogliastra Genetic Park - Talana	19247500
ORCADES HAP300	Orkney Complex Disease Study	18760389
ORCADES OMNIX	Orkney Complex Disease Study	18760389
Pegasus	Pegasus Study	
PIVUS	Prospective Investigation of the Vasculature in Uppsala Seniors	16141402
PMNS	Pune Maternal Nutrition Study	-
PREVEND	Prevention of REnal and Vascular ENd-stage Disease	11004219
PROSPER	Prospective Study of Pravastatin in the Elderly at Risk	12457784
QFS	Quebec Family Study	24533236
QIMR	Queensland Institure of Medical Research	
RAINE	Raine Study	
RS	Rotterdam Study	
SHIP	Study of Health in Pomerania	20167617
SHIP-TREND	Study of Health in Pomerania - Trend	20167617
SIGNET	Sea Islands Genetic Network	18835935
Sorbs	Sorbs Study	21559053; 22907691
SR	Silk Road Study	
STR	Swedish Twin Registry - Twingene study	23137839
THISEAS_CAD cases	The Hellenic study of interactions between SNPs & Eating in Atherosclerosis	20167083
THISEAS_CAD controls	The Hellenic study of interactions between SNPs & Eating in Atherosclerosis	20167083
TRAILS-population cohort	Tracking Adolescents' Individual Lives Survey	23021478
TwinGene	Swedish Twin Registry - Twingene study	23137839
TwinsUK_317K	Twins UK study	23088889
TwinsUK_610K	Twins UK study	23088889
ULSAM	Uppsala Longitudinal Study of Adult Men	1216390
VB	Val Borbera Study	19847309
WELLGEN	Wellcome Genetic Collection	-

WGHS	Women's Genome Health Study	18070814
YFS	Young Finns Study	18263651

Appendix 1b ROHgen Intra-Cohort Genotype QC

Cohort	Call rate [filter detail / N individuals excluded]	Heterozygosity [filter detail / N individuals excluded]	Ethnic outliers / other exclusions	Individuals post GT QC	Call rate [filter detail / N SNPs excluded]	SNP number in QC'd dataset
AEGS1	0.85	±3s.d.	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	657	0.97	403,832
AEGS2	0.98	±3s.d.	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	869	0.97	535,983
AGES-Reykjavik	0.97	NA	BeadStudio 0.4 Threshold;Gender mismatch; Mismatch previous genotypes	3219	97%/3,767	325,094
AIDHS/SDS	0.93	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS	1616	0.98	474,231
ALSPAC	0.98	FDR<1%	Ethnic outliers and cryptic relateds already removed	3326	0.97	526,688
AMISH	0.95	NONE	gender mismatch; pedigree mismatch	1827	0.95	318,792
ARIC	0.95	FDR < 1%	Ethnic outliers based on principal components; duplicates; gender mismatch; Discordant genotype with earlier TaqMan genotyping; 1 of each 1st degree relative pair	11562	0.9	685,812

Appendix 1b ROHgen Intra-cohort genotype QC

ASCOT-SC	0.99	separately <1%, >1% MAF, excl +/- 3 SD	Ancestry outliers by PCA; duplicates & related by 0.25 IBD pi_hat threshold; gender mismatch	2493		0.98	923,591
ASCOT-UK	0.95	NA	Ancestry outliers by PCA: iteratively removing individuals +/-6sd on first 10 PCs; duplicates, 1st & 2nd deg rels excluded	3804		0.97	283,291
ASPS	0.98	FDR<1% / 1	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	829		0.98	566,930
ASPS-Fam	0.98	FDR<1% / 0	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	388		0.98	617,191
B58C	0.97	No exclusions applied	Exclusion of duplicates, contamination, non-European identity	6491	None		519,040
BASEII	0.9	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1990		0.98	867,143
BBJ	0.98	NA	Ethnic outliers; related subjects	30322		0.99	477,784
BRIGHT	0.97	Het > 0.3 or Het < 0.225	Non European ancestry; duplicates; 1st or 2nd deg relatives; gender mismatch	1948		0.95	446,472
CARDIA_CARe_AA	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	867		0.95	739,824
CARL	0.97	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	630		0.97	309,430
CHOP - African American	0.95	NA	Ethnic outliers; duplicates; gender mismatch	1704		0.95	528,421
CHOP - Caucasian	0.95	NA	Ethnic outliers; duplicates; gender mismatch	2263		0.95	513,420

Appendix 1b ROHgen Intra-cohort genotype QC

CHS (African American)	0.95	NA	Gender mismatch, discordance with known sex or prior genotyping, lack of consent for DNA studies.	823	0.97	963,248
CHS (European)	0.95	NA	Gender mismatch, discordance with known sex or prior genotyping, lack of consent for DNA studies.	3271	0.97	306,655
CIDR_T2D	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	1808	0.95	741,072
CILENTO	0.95	NA	-	1512	0.95	189,751
CLHNS	0.97	None excluded	Duplicates; one member of 1st-degree relative pairs	1798	0.9	422,494
COLAUS	0.9	NA	-	5636	0.7	460,885
COPSAC2000	0.98	mean–(4*SD)	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	555	0.98	486,373
Corogene case	0.95	?	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	2235	0.95	554,987
Corogene control	0.95	?	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	1887	0.95	554,987
CROATIA_Korcula	0.97	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	897	0.98	307,625
CROATIA_Split _370CNV	0.97	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	499	0.98	321,456
CROATIA_Split+OMNIX +	0.98	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	491	0.98	679,002
CROATIA_Vis	0.97	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	960	0.98	289,827
DESIR	0.9	FDR<1%	Ethnic outliers; no duplicates	697	0.95	309,126
DNBC	0.98	3 sd	Ethnic outliers; duplicates; gender	2277	0.98	525,129

Appendix 1b ROHgen Intra-cohort genotype QC

			mismatch; IBS incompatible with pedigree			
EGCUT_370	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	2392	0.95	335,036
EGCUT_OMNI	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	7295	0.95	710,831
eMERGE_PAD	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	3048	0.95	520,994
EPIC	0.94	<23% or >30%: n= 20	Ethnic outliers; duplicates, crytptic relatedness	2417	0.9	397,438
ERF	0.98	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	2856	0.95	315,663
FamHS Human 1M-Duov3	0.98	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1490	0.98	921,414
FamHS Human 6100-Quadv1	0.97	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1674	0.98	536,610
FamHS HumHap550K	0.98	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	971	0.98	504,843
Fenland	0.95	27.35% or >28.82%	relatedness check, duplicate check	1402	0.9	362,055
FHS	0.97	5SD from mean	Ethnic outliers; excess Mendelian errors	8471	0.97	382,077
FINRISK/ENGAGE	0.95	NA	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	5312	0.95	257,671
FTC_1	0.95	excluded if F>0.05	Duplicates ; Gender mismatch; Sequenom fingerprint; MDS plot	964	0.95	511,568

Appendix 1b ROHgen Intra-cohort genotype QC

FTC_2	0.95	excluded if F>0.05 or F<-0.03	Duplicates; Gender mismatch; Sequenom fingerprint; MDS plot	1308	0.95	549,060
FTC_3	0.95	excluded if ±3SD from mean	Duplicates; Gender mismatch; Sequenom fingerprint	1780	0.95	259,726
FUSION	0.85	NA	ethnic outliers; duplicates; gender mismatch; T2D case sample; overlapping METSIM samples	1158	0.95	315,635
FVG	0.97	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1230 (OmniEx) + 330 (370k)	0.97	633354 (OmniEx) ; 330151 (370k)
GeneSTAR AA	0.9	NA	gender mismatch [n=2]; duplicate [n=1]; Mendelian error rate > 5% [n=4]; ethnic outliers [n=12]	1203	0.99	818,154
GeneSTAR EA	0.9	NA	gender mismatch [n=1]; duplicate [n=0]; Mendelian error rate > 5% [n=10]; ethnic outliers [n=7]	1988	0.99	817,738
GENOA (African American)	0.95	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1263	0.95	762,766
GENOA (European)	0.95	NA	Duplicates; gender mismatch; IBS incompatible with pedigree	1386	0.95	668,293
GoDARTS	0.98	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	2903	0.97	601,463
GOLDN	0.96	NA	4 duplicate samples deleted; Identified sex mismatches were due to sample switches that we were able to identify and correct using our family data	822	0.96	654,753
GOYA	0.95	35 individuals excluded	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	792	0.95	545,349

Appendix 1b ROHgen Intra-cohort genotype QC

GRAPHIC	0.9	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1012	0.97	616,550
GS:SFHS	0.98	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	9863	0.98	615,104
H2000/Genmets case	0.95	?	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	855	0.95	555,388
H2000/Genmets control	0.95	?	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	867	0.95	555,388
HBCS	0.95	?	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	1676	0.95	546,814
HELIC_MANOLIS	0.98	visual	Ethnic outliers; duplciates; sex discrepancies; GWAS concordance for exomechip, MAC=1 outliers for exomechip	1267	OmniExpress = MAF<5% 99% & MAF≥5% 95%; Exomechip = 95% Gencall & 99% zCall	719,305
HELIC_POMAK	0.98	visual	Ethnic outliers; duplciates; sex discrepancies; GWAS concordance for exomechip, MAC=1 outliers for exomechip	1007	OmniExpress = MAF<5% 99% & MAF≥5% 95%; Exomechip = 95% Gencall & 99% zCall	699,817
HPFS-Affymetrix	90-98%	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	3598	0.97	668,283
HPFS-Illumina	95-98%	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1368	0.97	459,999

Appendix 1b ROHgen Intra-cohort genotype QC

HPFS-Omni	95-98%	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1812	0.97	565,810
HRS	0.98	None	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	9991	0.98	1,029,958
HRS	0.98	None	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	2279	0.98	453,277
HUFS	0.9	NA	Ethnic outliers; duplicates; pedigree inconsistencies; relatives	1018	0.95	808,465
Hutterites	0.95	NA	IBS mismatch with pedigree; gender mismatch	1415	0.95	271,486
HyperGEN - African Americans	0.98		17 subjects excluded for poor quality DNA, suspected sample switches, gender mismatch, or Mendelian incompatibilities	1083	0.97	837,134
HyperGEN - Caucasians	0.98		49 subjects excluded for poor quality DNA, suspected sample switches, gender mismatch, or Mendelian incompatibilities	1270	0.97	358,327
HYPERGENES - Normotensives	0.95	Het > +3sd or Het < -3sd	Ethnic outliers; heterozigousity; duplicates; gender mismatch; IBS incompatible with pedigree; call rate < 95%	1709	0.97	840,212
InCHIANTI	0.98	NA	N/A	1210	0.98	495,343
INCIPE	0.98	Het > +3sd or Het < -3sd	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	940	0.97	651,801
INCIPE2	0.98	Het > +3sd or Het < -3sd	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1996	0.97	280,746

Appendix 1b ROHgen Intra-cohort genotype QC

INDICO_case	0.99	Het > +3sd or Het < -3sd	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1110	0.95	520,244
INDICO_control	0.99	Het > +3sd or Het < -3sd	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1105	0.95	519,022
IPM-BioMe	0.98	FDR<1%	Ethnic outliers; duplicates; gender mismatch; related individuals (PI_HAT > 0.2)	10511	0.99	575,981 (AA), 543,387 (EA), 574,377 (HA)
JHS	0.95	FDR<1%	Ethnic outliers; duplicates; gender mismatch.	3028	0.95	868,969
KORA F3	0.97	5sd	European descent	3077	0.98	2,380,310
KORA F4	0.97	5sd	mismatch of phenotypic and genetic gender; check for European ancestry; check for population outlier; mismatch with genotypes of same individual on other genotyping chips if available; individuals from KORA S4 cohort were removed	2927	0.98	558,446
LBC1921	0.97	NA	Ethnic outliers; gender mismatch, relatedness	517	0.98	549,692
LBC1936	0.97	NA	Ethnic outliers; gender mismatch, relatedness	1005	0.98	549,692
LHS_EA	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	1394	0.95	528,696
LOLIPOP_EW610	0.95	NA	Duplicates, gender discrepancy, contaminated samples, relatedness, call	927	0.95	544,620

Appendix 1b ROHgen Intra-cohort genotype QC

			rate <95%			
LOLIPOP_EWA	0.95	NA	Duplicates, contaminated samples, relatedness, samples already in EW610, call rate <95%	582	0.95	374,773
LOLIPOP_EWP	0.95	NA	Duplicates, contaminated samples, samples already in EW610, call rate <95%.	644	0.95	184,469
LOLIPOP_IA317	0.95	NA	Duplicates, samples already in IA610, gender discrepancy, ethnic outliers, contaminated samples, relatedness, call rate <95%.	2121	0.95	245,892
LOLIPOP_IA610	0.98	9 subj	Duplicates, gender discrepancy, ethnic outliers, contaminated samples, relatedness, call rate <95%.	6548	0.95	544,390
LOLIPOP_IAP	0.95	NA	Duplicates, contaminated samples, samples already in other IA data sets, call rate <95%.	638	0.95	170,055
LOLIPOP_OmniEE	0.99	26 subj	Duplicates, samples already in IA610 or IA317, gender discrepancy, ethnic outliers, contaminated samples, relatedness, call rate <98%, extreme heterozygosity	1018	0.98	692,266
MAYWOOD	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	743	0.95	859,332
MEGA FU	0.95	FDR<1%	Ethnic outliers, gender mismatch	1311	0.98	497,563
MESA	0.95	NA	Ethnic outliers; duplicates; gender mismatch;	6357	0.9	881,666
METSIM	0.85	NA	ethnic outliers; duplicates; gender mismatch; discordance with previous genotyping	10080	0.95	681,803
MICROS	0.98	NA	FDR<1%	1328	0.98	304,383
NHS-Affymetrix	90-98%	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with	4398	0.97	668,283

Appendix 1b ROHgen Intra-cohort genotype QC

			pedigree			
NHS-Illumina	95-98%	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	3493	0.97	459,999
NHS-Omni	95-98%	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	3364	0.97	565,810
NIGERIA	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype; cryptic relatedness	1188	0.95	794,766
NIHS	0.98	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	506	0.99	552,846
NSPHS_06	0.95	FDR < 1%	genetic outliers, twins, duplicates	691	0.9	306,086
NSPHS_09	0.95	FDR<1%	genetic outliers, twins, duplicates	345	0.9	631,503
NTR	0.95	F<-.10 & F > .10 / 61 subjects excluded	Ethnic outliers (detected with the help of 1000 Genomes PCs and parental birthplace; N=321); ; individuals with lesser genotyping quality detected with PCA; duplicates; gender mismatch; IBS incompatible with pedigree; CQC < 0.4 (a quality metric from Affymetrix representing how well allele intensities separate into clusters)	8815	0.95	498,592
OBA	0.9	FDR<1%	Ethnic outliers; no duplicates	664	0.95	317,054
OGP	0.97	FDR<1%	-	378	0.95	347,878
OGP Talana	0.97	FDR<1%	-	783	0.95	342,202
ORCADES HAP300	0.98	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	854	0.97	287,208
ORCADES OMNIX	0.98	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	1361	0.97	615,658

Appendix 1b ROHgen Intra-cohort genotype QC

Pegasus	94% (n=323)	< 16% or > 21% (n=7)	Ethnic outliers (<80% European ancestry); unexpected duplicates; cryptic relatedness	7,473 (4,561 prostate cancer cases + 2,912 controls)	0.95	1,240,751
PIVUS	0.95	3SD of mean	Ethnic outliers; duplicates; gender mismatch	949	95 % (99 % if MAF < 5 %)	738,583
PMNS	0.95	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree; PCA outlier; lack of phenotype	1038	0.95	625,046
PREVEND	0.95	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree; sample mixups	3649	0.95	259,583
PROSPER	97.5%	>3sd, 11 samples excluded	sex mismatch, duplicates, non-caucasion, familiar relationships	5244	0.975	557,192
QFS	0.95	NA	-	928	0.95	543,713
QIMR	0.95	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	19360 (12655 with at least RoH phenotype)	0.95	268,314
RAINE	0.97	HET < 0.30 / 3 individuals	Replicates, Relatedness (Pi > 0.1875), gender mismatch	1494	0.95	535,632
RS	0.98	FDR<0.1%	Ethnic outliers; duplicates; gender mismatch; familial relations	6291	0.98	520,025
SHIP	0.92	NA	duplicates; gender mismatch	4079	0.97	589,871
SHIP-TREND	0.94	NA	duplicates; gender mismatch	986	0.97	1,265,745

Appendix 1b ROHgen Intra-cohort genotype QC

SIGNET	0.95	NA	Ethnic outliers; duplicates; gender mismatch; incomplete phenotype	1303	0.95	817,797
Sorbs	0.94	NA	Ethnic outliers; duplicates; gender mismatch; IBS>0.2	705	0.95	378,513
SR	0.97	FDR<1%	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	664	0.97	612,292
STR	0.97	NA	1) Sex-check (heterozygosity of X-chomosomes); 2) Deviations in heterozygosity of more then 5 SD from the population mean; 3) Cryptic relatedness check	9617	0.97	581,537
THISEAS_CAD cases	0.95	>3SD	Ethnic outliers; duplicates; gender mismatch	359	0.98	733,202
THISEAS_CAD controls	0.95	>3SD	Ethnic outliers; duplicates; gender mismatch	543	0.98	733,202
TRAILS-population cohort	0.95	heterozygosity >4SD	duplicates; gender mismatch; non-caucasian	1354	0.95	255,254
TwinGene	0.98	NA	cryptic relatedness, gender mismatch	10728	0.97	644,556
TwinsUK_317K	0.98	heterozygosity across all SNPs ≥2 s.d.	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	2040	97% (SNPs with MAF.5%) or 99% (for 1% MAF < 5%)	290,787
TwinsUK_610K	0.98	heterozygosity across all SNPs ≥2 s.d.	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree	3614	97% (SNPs with MAF.5%) or 99% (for 1% MAF < 5%)	489,586
ULSAM	0.95	3SD of mean	Ethnic outliers; duplicates; gender mismatch	1116	95 % (99 % if MAF < 5 %)	1,621,833
VB	0.9	NA	Ethnic outliers; duplicates; gender mismatch; Mendelian errors	1785	0.9	332,887

Appendix 1b ROHgen Intra-cohort genotype QC

WELLGEN	0.95	NA	Ethnic outliers; duplicates; gender mismatch; IBS incompatible with pedigree; PCA outlier	1062	0.95	623,816
WGHS	0.98	NA	Consistency between self-report and multi-dimensional scaling using 1443 ancestry informative markers	23294	0.98	339,596
YFS	0.95	NA	Duplicate/related; Gender mismatch; Cryptic relatedness; Heterozygosity	2442	0.95	546,674

Appendix 1b ROHgen Intra-cohort genotype QC

Appendix 2 Copy of Manuscript: Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies

Joshi PK, Prendergast J, Fraser RM, Huffman JE, Vitart V, et al. (2013) Local exome sequences facilitate imputation of less common variants and increase power of genome wide association studies. PLoS One 8: e68604.

PLOS | ONE

Local Exome Sequences Facilitate Imputation of Less Common Variants and Increase Power of Genome Wide Association Studies

Peter K. Joshi[1], James Prendergast[2], Ross M. Fraser[1], Jennifer E. Huffman[2], Veronique Vitart[2], Caroline Hayward[2], Ruth McQuillan[1], Dominik Glodzik[1,2], Ozren Polašek[3,4], Nicholas D. Hastie[2], Igor Rudan[1], Harry Campbell[1], Alan F. Wright[2], Chris S. Haley[2,5], James F. Wilson[1,2]*, Pau Navarro[2]*

[1] Centre for Population Health Sciences, University of Edinburgh, Edinburgh, Scotland, United Kingdom, [2] MRC Human Genetics Unit, University of Edinburgh, Edinburgh, Scotland, United Kingdom, [3] Department of Public Health, University of Split, Split, Croatia, [4] Centre for Global Health, University of Split, Split, Croatia, [5] Roslin Institute, University of Edinburgh, Scotland, United Kingdom

Abstract

The analysis of less common variants in genome-wide association studies promises to elucidate complex trait genetics but is hampered by low power to reliably detect association. We show that addition of population-specific exome sequence data to global reference data allows more accurate imputation, particularly of less common SNPs (minor allele frequency 1–10%) in two very different European populations. The imputation improvement corresponds to an increase in effective sample size of 28–38%, for SNPs with a minor allele frequency in the range 1–3%.

Citation: Joshi PK, Prendergast J, Fraser RM, Huffman JE, Vitart V, et al. (2013) Local Exome Sequences Facilitate Imputation of Less Common Variants and Increase Power of Genome Wide Association Studies. PLoS ONE 8(7): e68604. doi:10.1371/journal.pone.0068604

Editor: Struan Frederick Airth Grant, The Children's Hospital of Philadelphia, United States of America

Received January 23, 2013; **Accepted** May 30, 2013; **Published** July 16, 2013

Funding: This research was supported by the Chief Scientist Office of the Scottish Government, the Medical Research Council, the European Union Framework Programme 6, the MS Society UK, the Volant Trust, the Royal Society, and the Croatian Ministry of Science Education and Sport. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pau.navarro@ed.ac.uk (PN); jim.wilson@ed.ac.uk (JFW)

These authors contributed equally to this work.

Introduction

Genome-wide association study (GWAS) meta-analyses routinely use genotype imputation [1]. Accurate imputation of less common variants (minor allele frequency MAF, 1–10%) may be particularly useful as commercial genotyping arrays often provide poor coverage of such variants, and imputation improves association power most for less frequent causal variants [2].

The recently released 1000 Genomes haplotypes [3] are a particularly large and dense reference panel that will be commonly used as an imputation reference panel, particularly in GWAS consortia. At the same time, theoretical studies and empirical studies using other primary reference panels, have shown that imputation accuracy in a study population can be increased by use of an additional reference panel such as whole genome or exome sequence data drawn from a subset of the population under study [2] [4] [5] [6] [7] [8] [9].

It is therefore useful to quantify the likely benefit of adding local reference data to 1000 Genomes data, particularly for less common variants, and especially if the population is genetically distant from the 1000 Genomes populations.

We used data from the CROATIA-Korcula and Orkney Complex Disease studies (ORCADES) [10] [11]. Both studies are family-based, cross-sectional community studies of the genetics of complex traits. The Croatian island of Korčula is in the Adriatic and the ORCADES study is based in the Orkney Isles in Scotland.

Genotypes obtained from the whole exome sequencing of 91/89 CROATIA-Korcula/ORCADES quality controlled samples were used to supplement the 1000 Genomes reference panel. We focused on less common (MAF 1–10%) exonic variants already in 1000 Genomes which, unlike low frequency, and rare (MAF<1%) or private variants, can be meta-analyzed in typically sized consortia.

In this paper, we therefore seek to determine if imputation accuracy can be improved by the addition of local sequences to a global reference panel.

Methods

The ORCADES and CROATIA-Korcula studies both had ethical approval for genetic research into the basis of complex traits, approved by the appropriate committees in each country. For ORCADES the committees were the Orkney Local Research Committee and the North of Scotland Research Ethics Committee (approval Orkney: 27/2/04). For CROATIA-Korcula the committees were the Ethics Committee of the Medical School, University of Split (approval id 2181-198-03-04/10-11-0008) and the NHS Lothian (South East Scotland Research Ethics Committees; REC reference 11/AL/0222). All participants provided written informed consent.

Array genotypes were obtained from Illumina Hap370CNV array, at 319,552 SNPs for CROATIA-Korcula subjects and Illumina Omni1 array at 1,140,419 SNPs or the Illumina Human

Hap300 array at 293,687 SNPs for ORCADES subjects. For ORCADES a common panel of intersecting Hap300 and Omni1 SNPs was first created. The panel for CROATIA-Korcula was then restricted to these SNPs, to ensure similar panel sizes.

Subjects to be sequenced were selected from the wider study populations that were genotyped on the Illumina Hap (370CNV/300) arrays to minimize relatedness, and thus to maximize representation of study population haplotypes. The selection was carried out by tracking the identity-by-descent sharing structure, as determined by the array genotypes using the program ANCHAP [12]. Whole exome sequences of 99/95 CROATIA-Korcula/ORCADES subjects were generated using the Agilent SureSelect All Exon 50 Mb kit and 234,746/217,015 variants were identified.

Quality control (QC) of genotyping array data, that were subsequently used for imputation, was in accordance with best practice for association studies [13] and is described in detail in Methods S1. As illustrated in Figure 1, post QC array data of 170,134/171,749 SNPs for 892/1158 Korčulan/Orcadian subjects were then pre-phased simultaneously (within each population) using SHAPEIT v1.r416 [14] [15] including the maximal pedigree structure permitted by the software (non-overlapping nuclear families) to create a phased set of study genotypes ready for imputation using IMPUTE2 v2.2.2 [16]. The simultaneous phasing of all (892/1158) study subjects allowed all these subjects' phasing to inform the phase of the ~100 subjects taken forward as a reference panel and for imputation.

Exome sequence data were also subjected to rigorous QC to ensure they were of high quality so that that the local reference panel we created did not have a significant number of incorrect haplotypes. Variants were called by first aligning the raw sequence data to the human hg19 reference genome using the Stampy short read aligner [17] (with BWA utilized as a pre-mapper [18]). Genotype calls were produced from the resulting alignments using GATK's unified genotyper, following GATK's recommended best practice for variant detection from exome sequence datasets [19]. Variants were required to have a phred-scaled quality of at least 40. Individual sample genotype calls with a phred-scaled quality less than 20 were regarded as missing. Variants that were called in less than 50% of subjects, or with a minor allele frequency of less than 0.75% were removed (hence inclusion required at least two minor alleles across samples). All variants that mapped to more than one homologous region or failed a test of Hardy-Weinberg equilibrium (HWE) with a p-value of less than 10^{-4}, were also removed, leaving 99/95 CROATIA-Korcula/ORCADES subjects genotyped for 102,192/97,052 variants. The HWE test was a more stringent test than for the array data reflecting lower sample numbers and the desire to particularly ensure integrity for reference data. We restricted our analysis to individuals with exome sequences and merged the exomes with the array data for these subjects. Subjects/variants with more than 50/30 mismatching calls, between the array and sequence data were excluded, although no variants failed this test. This resulted in exomes for 93/90 subjects genotyped at 102,192/97,052 exonic SNPs being merged with array data at 170,134/171,749 SNPs for these individuals. The resulting panels had 265,929/262,513 variants which were 99.91%/99.92% concordant, based on the genotypes called on both panels for 6,397/6,285 overlapping variants. As the overall genotypic concordance could mask discrepancies for minor alleles, particularly the less common variants of interest, concordance rates for minor allele calls were calculated in the MAF 1–3% range separately. Only 1/1 (CROATIA-Korcula/ORCADES) call was discrepant on each overlapping panel, giving minor allele concordance of 99.7% in both studies for these variants.

8,150/10,964 Korčulan/Orcadian variants other than single base substitutions, for example insertions or deletions, were excluded. 119/110 conflicting map positions and individuals called at fewer than 80% of the combined SNP panel were then excluded, leaving 91/89 subjects typed across 257,633/251,439 SNPs. Our focus was on the potential to improve power in meta-analyses, so polymorphisms that were unique to each cohort were excluded. This was done by comparison to the 1000 Genomes project map and those variants not present in the 1000 Genomes reference data or with mismatches in allele codes were excluded.

The merged sequence and array data consisting of 233,195/232,096 variants for 91/89 subjects were then phased by SHAPEIT, using the recommended N_e of 11,418 and the default settings [14], to create reference haplotypes, as shown in the lower half of Figure 1.

Having created suitable post-QC array data and secondary reference panels, imputations were performed using genome-wide array data plus (i) 1000 Genomes haplotypes [2] alone or (ii) 1000 Genomes haplotypes together with local data as reference panels. Both imputations were then compared with known genotypes and an assessment of accuracy across all subjects was made for each SNP, as illustrated in Figure 2.

Imputation of the 91/89 subjects with and without the benefit of local reference data was carried out using IMPUTE2, using the phased reference panel option, the phased array data haplotype option, and with the software splitting the genome into chunks, which had been predetermined to be less than 5 Mb in size and avoiding crossing the centromeres. N_e was set to 20,000; all other settings were left at their default values. For the one panel

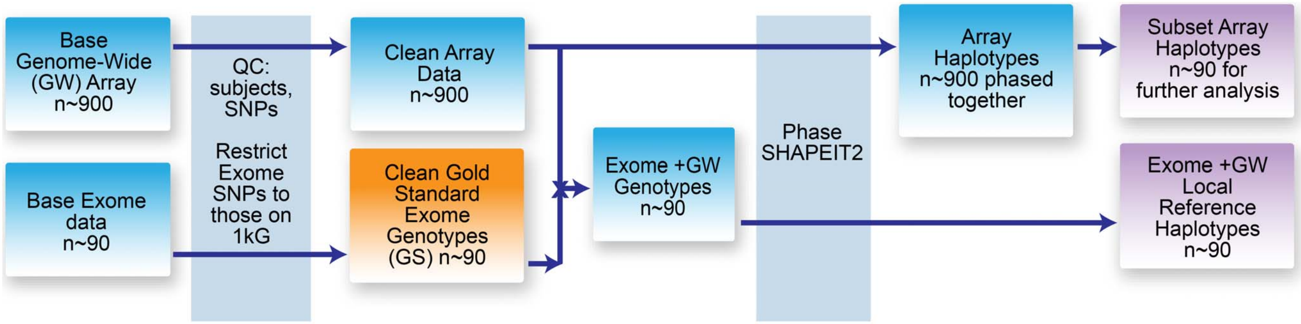

Base Genome-Wide (GW) Array n~900
Base Exome data n~90
QC: subjects, SNPs
Restrict Exome SNPs to those on 1kG
Clean Array Data n~900
Clean Gold Standard Exome Genotypes (GS) n~90
Exome +GW Genotypes n~90
Phase SHAPEIT2
Array Haplotypes n~900 phased together
Subset Array Haplotypes n~90 for further analysis
Exome +GW Local Reference Haplotypes n~90

Figure 1. Preparation of array data and local reference panel for imputation. The genotype data were quality controlled and phased. These data were then used in further downstream analysis.
doi:10.1371/journal.pone.0068604.g001

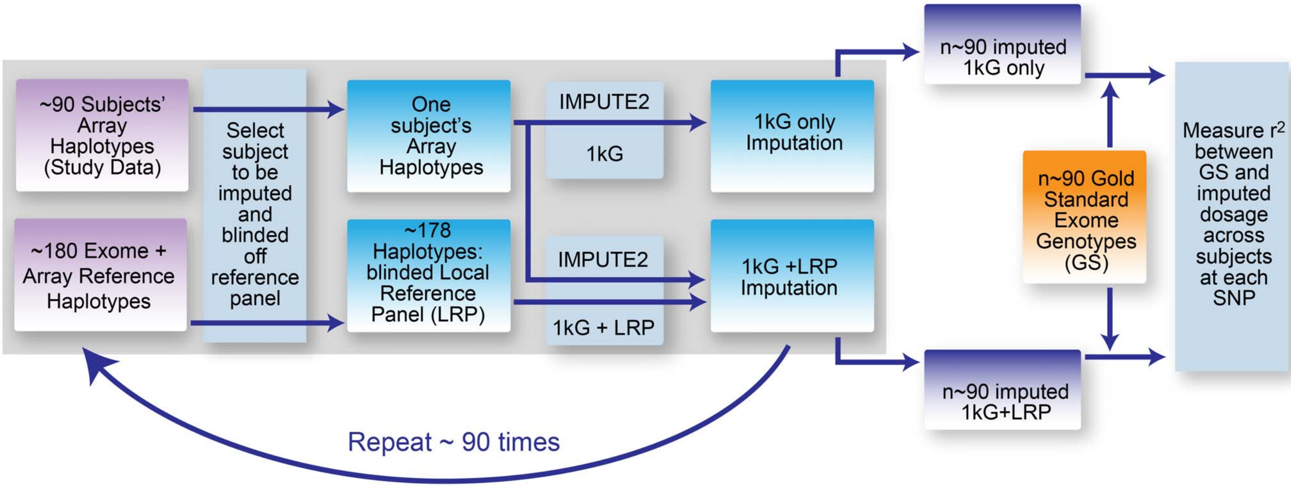

~90 Subjects' Array Haplotypes (Study Data)
~180 Exome + Array Reference Haplotypes
Select subject to be imputed and blinded off reference panel
One subject's Array Haplotypes
~178 Haplotypes: blinded Local Reference Panel (LRP)
IMPUTE2
1kG
IMPUTE2
1kG + LRP
1kG only Imputation
1kG +LRP Imputation
n~90 imputed 1kG only
n~90 Gold Standard Exome Genotypes (GS)
n~90 imputed 1kG+LRP
Measure r2 between GS and imputed dosage across subjects at each SNP
Repeat ~ 90 times

Figure 2. Illustration of the procedure to estimate imputation accuracy. We used a drop one-out crossvalidation approach. For the imputation step each subject was removed from the reference panel in turn, and this subject's exome sequence SNPs were then imputed using either the 1000 Genomes reference panel alone or in conjunction with a second local reference panel. All subjects' imputed allelic dosages were then compared with the exome sequence genotype data ("gold standard").
doi:10.1371/journal.pone.0068604.g002

imputation, the 1000 Genomes Phase 1 worldwide integrated variant set (March 2012 release) [3] as available on the IMPUTE2 website [16] was used. The two-panel imputation added the phased local reference data as a secondary panel (we did not use the merge panels option). All other settings for the two-panel imputation we were identical to the one panel imputation. We performed imputations for each subject with local exome data separately, with the study subject's own haplotypes removed from the secondary reference panel so that the haplotypes of the individual to be imputed were not present in the reference data. For a given SNP, the accuracy (r^2) of the allelic dosages imputed was measured across samples against the known exome sequence-called genotypes.

As evidenced by the genome-wide SNP array concordance data, noted above, there was close agreement between the exome sequence and independent genotyping data, indicating that the sequences were a suitable gold standard. Furthermore exome array data were also available for the CROATIA-Korcula study (although not ORCADES) and concordance between exome array and exome sequence genotypes was 99.5% and was similar across all MAF bands.

The dual use of exome sequences both as a secondary reference panel and as the gold standard to obtain imputation accuracy was considered appropriate since a subject's imputation panel did not include their own sequence, avoiding circularity at the imputation stage.

Results

We found a significant increase in accuracy (r^2 of imputed against known allele dosages across samples for a given SNP) from use of a local reference panel, which was often substantial for less common variants (Table 1).

Variants with a minor allele frequency in the range 0.01–0.032 showed an increase in imputation accuracy of 0.193/0.167 (38%/28% improvement) for CROATIA-Korcula/ORCADES and 0.112/0.089 (15%/11% improvement) for variants with MAF between 0.032 and 0.100. The high accuracy of the 1000 Genomes imputation for more common variants (MAF >0.1) provided more limited scope for improvement in this category, although even for the most common variants (MAF>0.32) the accuracy of imputation increased by 0.039/0.031 (4%/3% improvement) for CROATIA-Korcula/ORCADES after adding the second (local) reference panel.

Much of the improvements arise from SNPs that have an r^2 close to zero with the 1000 Genomes-only imputation and which were imputed more accurately with the addition of the local panel (Figure 3). For CROATIA-Korcula/ORCADES 12%/9% of all SNPs imputed poorly (r^2<0.2) using 1000 Genomes data alone. About one-fifth (17.1%/19.9%) of these poorly imputed SNPs imputed well (r^2>0.8) after the addition of the local reference panel.

SNPs that were less frequent in 1000 Genomes than in our sequences generally improved more, as illustrated in Figure 4, where areas of greater improvement are generally observed towards the right-hand side in the figure. The effect is more pronounced in Korčula and is particularly marked for variants where MAF is less than 1% on 1000 Genomes European panel.

Counts of the SNPs in each cell of Figure 4 are shown in table S1.

We also looked at r^2 increase as a function of European 1000 Genomes MAF. As stated above, for SNPs with a MAF of 1–3.2% in our local sequences, the mean increase in r^2 was 0.193/0.167. For these SNPs, the increase in r^2 was 0.297/0.264 for those in the European 1000 Genomes MAF band <1%, 0.137/0.112 for MAF band 1–3.2% and 0.086/0.072 for MAF >3.2%.

Discussion

Our results show that use of a secondary local reference panel in addition to the 1000 Genomes reference haplotype data can significantly increase the quality of imputations, particularly for less common alleles and the improvement is greater when the study population is genetically further from the populations in the reference data.

We estimated imputation accuracy using a leave-one-out cross-validation approach, in which we compared known genotypes to imputed ones using either the 1000 Genomes reference panel alone or accompanied by a panel obtained from sequence data of individuals from our study populations. Although we took care in

Table 1. Mean accuracy of imputation (r^2 of allelic dosage across all samples for a SNP) averaged across SNPs split by Minor Allele Frequency (MAF).

MAF	**1–3.2%**		**3.2–10%**		**10–32%**		**>32%**	
Population	**Korčula**	**Orkney**	**Korčula**	**Orkney**	**Korčula**	**Orkney**	**Korčula**	**Orkney**
N SNPs	12132	12123	11548	10677	16243	15262	10174	9265
r^2 1kG	0.504	0.586	0.729	0.778	0.868	0.894	0.894	0.913
r^2 1kG+LRP	0.697	0.753	0.841	0.867	0.916	0.931	0.934	0.944
Increase r^2	0.193	0.167	0.112	0.089	0.049	0.037	0.039	0.031
Std dev.	0.309	0.295	0.182	0.157	0.093	0.078	0.074	0.065
Inc. Sample	38%	28%	15%	11%	6%	4%	4%	1%

MAF bins increase by factors of $\sqrt{10}$, to create four exponentially increasing bins.
N SNPs: number of SNPs in MAF bin.
1kG: 1000 Genomes used as reference panel.
1kG+LRP: 1000 Genomes plus local reference panel.
Increase r^2: Average across all SNPs in MAF bin increase in r^2.
Std dev: The standard deviation (across SNPs) of the increase in r^2 at each SNP.
Inc. Sample: Increase in effective sample size for GWAS.
The standard errors of mean increases are less than 0.003. All improvements in r^2 are significantly different from zero and significantly different between MAF bands ($P<0.001$, two-sided t tests).
doi:10.1371/journal.pone.0068604.t001

our cross-validations to avoid circularity by using the leave-one-out approach in the imputations, for practical reasons, especially computing time, the phasing stage was done only once including all subjects (and therefore included the subject being blinded at the imputation stage). We acknowledge that this could potentially slightly inflate the reported increase in accuracy when using the second reference panel.

Imputation accuracy is not only affected by the quality and composition of the reference data used, but also by the design of the genotyping array, in particular array density and whether the array captures population specific variants [20]. A dense, locally relevant array used to genotype the study population will improve the quality of imputation compared to a less dense one, when using a global reference panel, and thus reduce the potential scope for improvement when adding local sequence data. However, where the study population's haplotypes are distinct, due to recombination, from the reference panel population, the use of a denser array can be expected to improve the imputation but the denser array will also allow even better matching of local haplotypes, and so there should be a further benefit from use of a local secondary reference panel.

Consistent with this hypothesis, the accuracy of base imputations using only the 1000 Genomes reference panel was greater for ORCADES than CROATIA-Korcula, presumably due to the greater proximity of Orkney to subjects in the 1000 Genomes reference panel. Twenty three Orcadians, 77 mainland British and

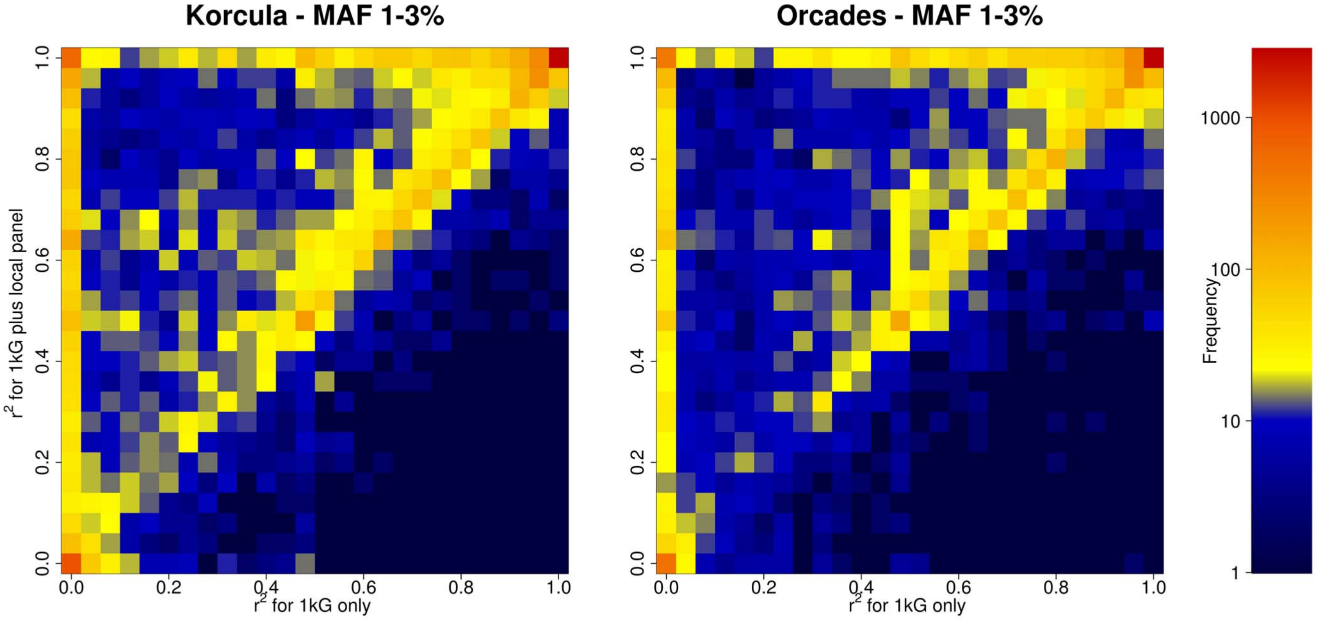


Figure 3. Frequency plot of imputation accuracy (r^2) using 1000 Genomes data alone against 1000 Genomes plus a local reference panel for SNPs with Minor Allele Frequencies (MAF) of 1–3.2%.
doi:10.1371/journal.pone.0068604.g003

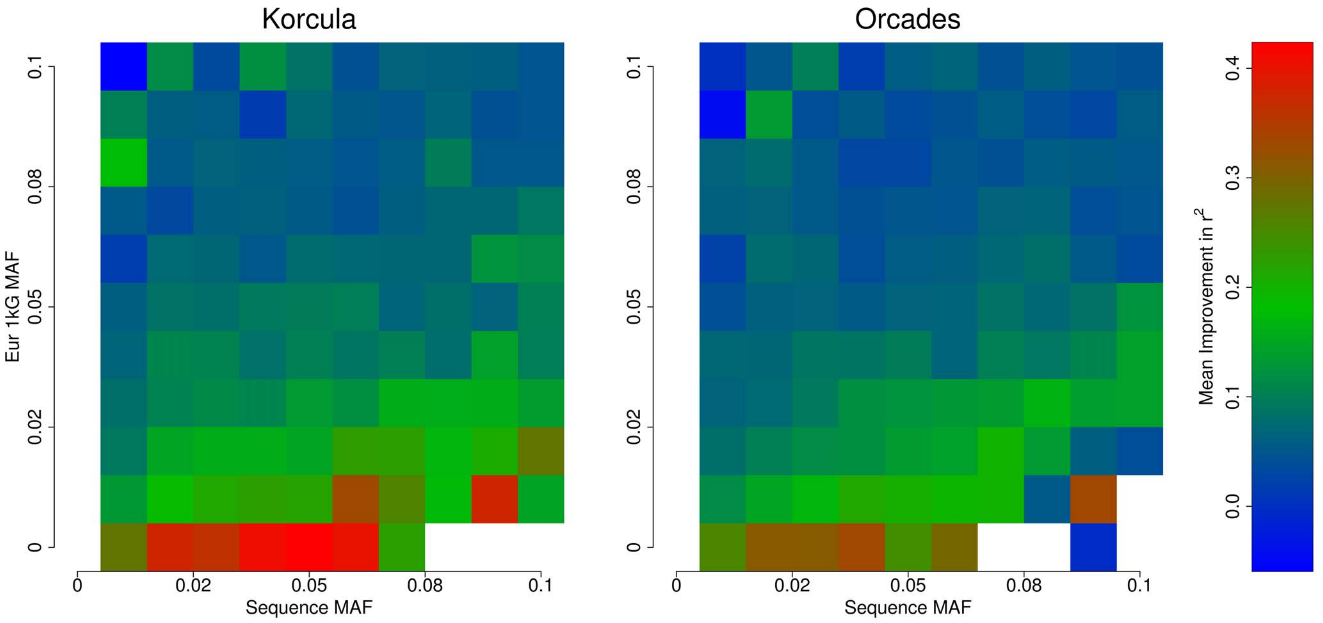


Figure 4. Plot of mean improvement in imputation accuracy (r^2) for SNPs with minor allele frequency (MAF) in the range 1–10% in our exome sequence data.
doi:10.1371/journal.pone.0068604.g004

100 of northern European ancestry individuals are present in the 1000 Genomes data, and principal component analysis shows that Balkan populations (such as Korčula) are more distant from the nearest subjects in 1000 Genomes (Tuscans, from central Italy, N = 100), than the variation observed within the British Isles [3] [21]. This suggests to us that, as might be expected, imputation improvement due to addition of local data will be most marked for populations genetically distant from 1000 Genomes samples. Whilst part of the benefit arises from including reference data with allele frequencies closer to the study population, the capture of representative local haplotypes further contributes to the increase in imputation accuracy, and this latter effect will be more marked, or at least require fewer local subject to be sequenced, in isolated populations, where fewer distinct haplotypes will be segregating.

Similarly the much greater improvements in accuracy for SNPs where the MAF is greater in our sequences than 1000 genomes, perhaps not surprisingly, shows that local sequences will add value to imputations in regions of the genome where drift, or other forces, have created a distinct genetic structure.

Comparing these results with those of other researchers who have examined the benefits of study specific reference panels, often using 1000 Genomes like us or HapMap [22] as primary panels, whilst illuminating, is not straightforward. Inevitably, different types and sizes of reference panels are used, as well as different genotyping arrays for the subjects whose genotypes are to be imputed. This is further complicated by different study protocols and differing genetic structure of the study populations. With these caveats, our results of an r^2 of 0.70–0.75 from 90 reference panel subjects in addition to 1000 Genomes seem consistent with those of Liu et al [9] and Auer et al [8], for MAF 1–3%. Neither of these studies used a global reference panel, but Liu et al, in their verification step, attained an r^2 of around 80% with ~2,000 subjects on their (array data) reference panel with unfiltered results, whilst Auer et al obtained an r^2 of 82% with 761 exome reference panel subjects, albeit filtering out lower quality results, using an Rsq threshold of 0.8, where Rsq is equivalent to the squared correlation between nearby imputed and genotyped SNPs [8]. Furthermore the latter study demonstrated that the use of exome imputation can reveal genome-wide significant associations, not discovered by conventional genotyping arrays, as did the study by Holm et al [23], who were able to discern a local rare variant causing sick sinus syndrome, in a large Icelandic study, due to the benefit of adding 87 whole genome sequences to the reference data for their imputation.

Many aspects of our study were similar to a study by Surakka et al [6]. Their Finnish study used 200 (CEU+TSI) HapMap [22] subjects as their primary reference panel and added 81 local subjects genotyped by a genome wide array. For alleles with a MAF <5%, they obtained a median r^2 of 90% for their global panel only imputation rising to 94% after the addition of their local panel. In our study, we report mean r^2, but our median r^2 was 0.77/0.83 rising to 0.88/0.92 after adding the local reference panel for CROATIA-Korcula/ORCADES for a MAF bucket 3–5%. The choice of a 3–5% MAF is intended to correspond to typical array SNPs with MAF<5%. Our results therefore appear consistent with the results of Surakka et al. despite the differences in study design. The study by Uricchio et al [7] obtained much higher mean r^2 (99%), and the technique used for imputation, identifying runs of identity-by-descent (IBD), should be particularly accurate, but its application is restricted to populations which share long haplotypes to a much greater extent than is common in most genetic studies, and we therefore feel our strategy of using 1000 Genomes reference data and adding sequence data from a subset of one's own study subjects is a good practical way forward for many studies.

A proportionate increase in r^2 has the same effect on power as a corresponding increase in study size [24] so the use of high quality sequence data has the potential to provide substantially greater power in GWAS studies for less common variants, particularly those very poorly imputed using 1000 Genomes alone but well imputed with the addition of local exome sequence data.

Our study focused on the exome, but the results should extend to any other genomic region of interest. Moreover, the similar results obtained in our study for two independent populations suggest that corresponding benefits will be found in other studies.

The meta-analysis of multiple populations imputed using local exome sequence data will likely identify new SNP associations. However the amount of variance explained by less common variants individually is likely to be small and will make their detection challenging. This will put increasing emphasis on the use of analytical methods that consider jointly groups of variants, be it gene [25], regional heritability [26] or network based analyses [27]. Such analyses can also incorporate the potentially valuable information provided by variants private to individual populations including the 24,438/19,343 variants identified by the exome sequencing of the CROATIA-Korcula and ORCADES samples that are not present in 1000 Genomes and hence we have not considered here.

Given the cost and significant practical difficulties in subject recruitment, sequencing a subset of cohort members, for either part or all of the genome, and using these results for imputation will provide significant added value to association studies.

Supporting Information

Methods S1 Quality Control of Array Data.
(DOCX)

Table S1 Counts of SNPs in each cell underpinning Figure 4.
(DOCX)

Acknowledgments

We thank the people of Orkney and Korčula, the data collection teams on both islands and the admin team in Edinburgh and Croatia. The SNP genotyping for both cohorts was performed in Helmholtz Zentrum München, Neuherberg, Germany.

Author Contributions

Conceived and designed the experiments: CSH JFW PN. Performed the experiments: PKJ. Analyzed the data: PKJ JP RFM. Contributed reagents/materials/analysis tools: JEH VV CH RM DG OP NDH IR HC AFW CSH JFW PN. Wrote the paper: PKJ CSH JFW PN.

References

1. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, et al. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17: R122–128.
2. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906–913.
3. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
4. Zeggini E (2011) Next-generation association studies for complex traits. Nat Genet 43: 287–288.
5. Jewett EM, Zawistowski M, Rosenberg NA, Zollner S (2012) A coalescent model for genotype imputation. Genetics 191: 1239–1255.
6. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, et al. (2010) Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res 20: 1344–1351.
7. Uricchio LH, Chong JX, Ross KD, Ober C, Nicolae DL (2012) Accurate imputation of rare and common variants in a founder population from a small number of sequenced individuals. Genet Epidemiol 36: 312–319.
8. Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, et al. (2012) Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am J Hum Genet 91: 794–808.
9. Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, et al. (2012) Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. Genet Epidemiol 36: 107–117.
10. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, et al. (2008) Runs of homozygosity in European populations. Am J Hum Genet 83: 359–372.
11. Polasek O, Marusic A, Rotim K, Hayward C, Vitart V, et al. (2009) Genome-wide association study of anthropometric traits in Korcula Island, Croatia. Croat Med J 50: 7–16.
12. Glodzik D, Navarro P, Vitart V, Hayward C, McQuillan R, et al. (2013) Inference of identity by descent in population isolates and optimal sequencing studies. Eur J Hum Genet.
13. Weale ME (2010) Quality control for genome-wide association studies. Methods Mol Biol 628: 341–372.
14. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9: 179–181.
15. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955–959.
16. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.
17. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936–939.
18. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
19. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.
20. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499–511.
21. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101.
22. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
23. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, et al. (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 43: 316–320.
24. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1–14.
25. Huang H, Chanda P, Alonso A, Bader JS, Arking DE (2011) Gene-based tests of association. PLoS Genet 7: e1002177.
26. Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2012) Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. PLoS One 7: e46501.
27. Cabrera CP, Navarro P, Huffman JE, Wright AF, Hayward C, et al. (2012) Uncovering networks from genome-wide association studies via circular genomic permutation. G3 (Bethesda) 2: 1067–1075.