THE UNIVERSITY *of* EDINBURGH

Homozygosity, inbreeding and health in European populations

Ruth McQuillan

PhD Thesis

The University of Edinburgh

2009

Table of Contents

Abstract

Inbreeding results in increased levels of homozygosity for deleterious recessive alleles, leading to increased incidence of monogenic disease in inbred families. It has also been suggested that inbreeding increases the risk of diseases such as cancer and heart disease, implying a role for the combined effects of many recessive alleles distributed across the genome. A better understanding of the links between inbreeding, homozygosity and disease is therefore of interest to those concerned with understanding the genetic architecture of complex disease. A homozygous genotype is defined as autozygous if both alleles originate from the same ancestor. Quantifying inbreeding involves quantifying autozygosity. A new, observational method of quantifying autozygosity using genomic data is developed here. Based on runs of homozygosity (ROH), this approach has a sound theoretical basis in the biological processes involved in inbreeding. It is also backed by strong empirical evidence, correlating strongly with pedigree-derived estimates of inbreeding and discriminating well between populations with different demographic histories. ROH are a signature of autozygosity, but not necessarily autozygosity of recent origin. Short ROH are shown to be abundant in demonstrably outbred individuals and it is suggested that this is a source of individual genetic variation which merits investigation as a disease risk factor, although denser genotype scans than those used in the present study are required for the reliable detection of very short ROH. In the absence of such dense scans, it is suggested that ROH longer than 1 or 1.5 Mb be used to estimate the effects of inbreeding on disease or quantitative physiological traits (QT), and that a simple measure of homozygosity be used to investigate overall recessive effects. Evidence for recessive effects on 13 QT important in

cardiovascular and metabolic disease was investigated in 5 European isolate populations, characterised by heightened levels of inbreeding. A significant decrease in height was associated both with increased homozygosity and (to a lesser extent) with increased ROH longer than 5 Mb (i.e. inbreeding) estimated using a 300,000 SNP panel. No evidence was found for recessive effects on any of the other QTs. Evidence for recessive effects on colorectal cancer risk were investigated in two outbred case control samples typed with a 500,000 SNP panel. Cases were significantly more homozygous and had more of their genome in short ROH than did controls. Cases were significantly more homozygous than controls even when inbred individuals were removed from the sample. There was also some evidence of an inbreeding effect, with inbred subjects having slightly significantly higher odds of colorectal cancer than outbred subjects. This study provides evidence of recessive effects on a common, complex disease in outbred populations and on height in both inbred and outbred populations and shows that such effects are not solely attributable to increased levels of homozygosity resulting from recent inbreeding. Individual variation among outbred individuals in the proportion of the genome that is homozygous may be important in disease risk. The development of denser genotype scans will facilitate better enumeration of short ROH in outbred individuals so that these can be properly enumerated and investigated as a disease risk factor.

Acknowledgements

This Thesis is the result of my involvement in the Orkney Complex Disease Study (ORCADES), an ongoing, family-based, cross-sectional study led by Dr James Wilson, that seeks to identify genetic factors influencing cardiovascular and other disease risk in the population of the North Isles of Orkney. I would like to begin by thanking the people of Orkney, without whose public-spirited willingness to volunteer for ORCADES this thesis would not have been possible. ORCADES data collection is undertaken by Lorraine Anderson and the research nurse team in Orkney, with administrative support provided by Kay Lindsay, Rosa Bisset and the administrative team at Public Health Sciences, University of Edinburgh. ORCADES is supported by the Chief Scientist Office of the Scottish Executive, the MRC Human Genetics Unit, the Royal Society, the Edinburgh Wellcome Trust Clinical Research Facility and the European Union Framework Program 6. Between September 2005 and August 2008, I was supported by a University of Edinburgh College of Medicine and Veterinary Medicine PhD Studentship.

I joined the project in 2005, as data collection was beginning in Orkney. From 2005 to 2006, my main role involved tracing the pedigrees of study participants using birth, marriage, death and census records held by the General Register Office for Scotland in Edinburgh. A substantial amount of this work was also done by Ruby McMenemy and Chris Franklin. This aspect of the project was made immeasurably easier by the help, advice and information provided by Sam Harcus and George Gray

of the Orkney Family History Society and by the staff of the Orkney Library and Archive.

On completion of the Orkney pedigree data base, which contains the records of over 12,000 individuals, I moved onto quality control and analysis of ORCADES genotypic data. Many people provided IT support during this period. Graeme Grimes at the MRC Human Genetics Unit (HGU) in particular was instrumental in getting large, unwieldy data files into usable formats. Sarfraz Mohammed, Dr Dave du Feu and Lesley McGoohan of Public Health Sciences, University of Edinburgh, also provided invaluable support. Dr Andrew MacLeod from the MRC HGU wrote the programme to produce figure 3.8. Craig Nicol, also from MRC HGU, improved the graphics quality of all the figures in chapter 3, bringing them up to the standard required for publication.

In September 2008, the material in chapter 3 of this thesis was published in the American Journal of Human Genetics (reproduced here in appendix 2). In addition to the ORCADES sample, this paper features data from a study in a Croatian genetically isolated population (CROAS) and from a Scottish colorectal cancer case-control study (SOCCS). I would like to thank Marijana Pericic, Lovorka Barac-Lauc, Nina Smolej-Narancic, Branka Janicijevic, Dr Ozren Polasek, Professor Pavao Rudan, Dr Caroline Hayward, Dr Veronique Vitart and Dr Igor Rudan from CROAS and Dr Albert Tenesa, Dr Susan Farrington, Dr Evropi Theodoratou, Professor Harry Campbell and Professor Malcolm Dunlop from SOCCS for making this collaboration possible. I would also like to thank Dr Anne-Louise Leutenegger from Inserm, Paris,

and Dr Albert Tenesa from MRC HGU for invaluable comments on the first draft of the published paper.

An important part of the published paper, and of chapter 3 of this thesis, is an analysis of copy number variation in the ORCADES sample. This work was undertaken entirely by Dr Rehab Abdel-Rahman, a fellow PhD student here. This is indicated in the thesis text.

ORCADES is part of the Eurospan Special Populations Research Network (EUROSPAN), a collaboration between 5 research programmes in genetically isolated populations throughout Europe led by Professor Harry Campbell. Chapter 5 of this thesis is an analysis of quantitative physiological traits in the five EUROSPAN populations led by Professor Cornelia van Duijn (ERF, the Netherlands), Professor Ulf Gyllensten (NSPHS, Sweden), Dr Igor Rudan (CROAS, Croatia) and Professor Peter Pramstaller (MICROS, Italy). Dr Colm O'Dushlaine and Chris Franklin undertook considerable work cleaning and formatting data so that all I had to do was analyse it. Professor Peter Visscher of the Queensland Institute of Medical Research, Australia, provided advice on interpreting the results of the analysis of height data. Dr Sarah Wild of Public Health Sciences provided advice on appropriate adjustments to blood pressure and lipid traits for those taking anti-hypertensive and lipid lowering medication. Dr Niall Anderson of Public Health Sciences provided statistical advice and support on a number of occasions, as did Mirna Kirin.

Chapter 6 is an analysis of colorectal cancer case-control data provided by 2 studies: SOCCS (see above) and a London-based study led by Professor Ian Tomlinson of the Institute of Cancer at St Bartholomew's Hospital.

I would like to thank my supervisors, Professor Harry Campbell, Dr Veronique Vitart and particularly, Dr James Wilson for their support, advice and encouragement throughout this project. Dr Wilson's energy, enthusiasm and ideas have been inspirational and it has been a great pleasure to work with him. More specifically, his suggestion that an analysis of runs of homozygosity might be an interesting way to compare individuals and populations with different ancestries has proven very fruitful. My colleague and fellow PhD student Chris Franklin also deserves special mention for his unfailing willingness to help, advise and explain on numerous occasions and a wide range of topics.

Finally, I would like to thank my wonderful family for their love and understanding over the last four years. My parents, Anne and Geoff McQuillan, have always encouraged me and taken an interest in my work. Anna, Oscar and Louis have put up with a distracted and disorganised mother with very good humour and without complaint. Finally, I cannot begin to list all the ways in which my husband Mark has supported me. In practical terms, he has allowed me time and space when I have needed it, shouldering more than his fair share of domestic responsibilities. Just as importantly, by his unfailing belief in me he has given me the confidence to see this through.

Declaration

I hereby declare that:

1) I composed this thesis entirely on my own;

2) The work presented in the thesis is my own, and for the elements of the thesis that were a result of my work within a group of scientists, I declare that I have made a substantial contribution to the work, and I have clearly indicated this contribution in the Acknowledgements;

3) This work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, May 2009.

Ruth McQuillan

Notes

Some of the material included in this thesis has been published before submission. This was done with the approval of my supervisors.

The publisher's formal permission was obtained for the copy of the published paper attached in **Appendix 2.**

Abbreviations

CRC – colorectal cancer

GWAS – genomewide association study

HWE – Hardy Weinberg equilibrium

MAF – minor allele frequency

OR – odds ratio

QC – quality control

QT – quantitative trait

ROH – run of homozygosity

SES – socio-economic status

SNP – single nucleotide polymorphism

Chapter 1: Introduction

1.1 Key concepts

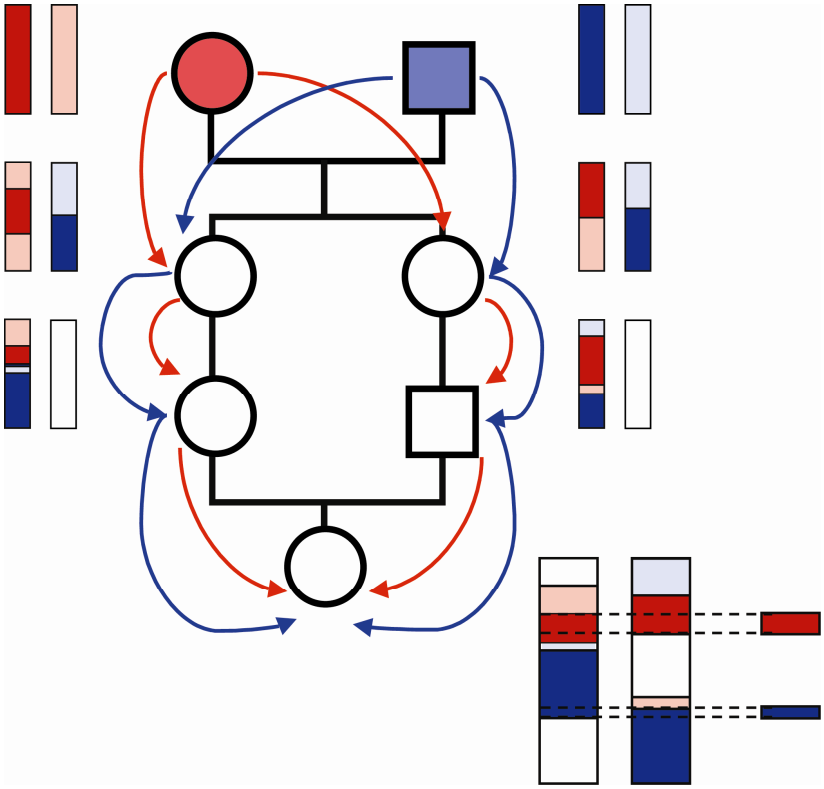
This thesis explores the population genetic concepts of inbreeding, autozygosity and homozygosity and their relation to human health in European populations.

Homozygosity is the state of having two identical alleles or genomic segments at corresponding loci on homologous chromosomes. Homozygosity is often partitioned into **autozygosity**, where the alleles or genomic segments are identical because they are inherited from a common ancestor, and "chance homozygosity", or **allozygosity**, where the two alleles have the same DNA sequence but are not inherited from a common ancestor (Purcell, Neale et al. 2007). The terms are not precisely defined, with autozygosity usually implying inheritance from a recent common ancestor only a few generations back and "chance homozygosity" being used to describe all homozygosity not derived in this way, including short homozygous segments inherited from very ancient common ancestors.

In addition, there are two related terms, which refer to the sharing of identical alleles among individuals or gametes, with or without homozygosity. Two alleles at a locus are said to be **identical-by-state** (IBS) if they have the same DNA sequence, regardless of whether they are inherited from a common ancestor. Two alleles at a locus are said to be **identical-by-descent** (IBD) if they are identical because they are inherited from a common ancestor.

Inbreeding, or mating between individuals who share a common ancestor, increases homozygosity because it increases the probability that at any given locus both alleles will be inherited identical-by-descent (IBD) from a common maternal and paternal ancestor. This is shown graphically in Figure 1.1. This probability can be calculated and expressed as the inbreeding coefficient, F (denoted throughout this thesis F_{ped} to distinguish it from genomically derived measures of inbreeding) which is defined as the probability of inheriting two alleles IBD at an autosomal locus, or equivalently, the average proportion of the autosomal genome inherited IBD (Hartl and Clark 1997).

Figure 1.1: Pedigree of the offspring of first cousins



An example chromosome is illustrated. The female common ancestor is red. The chromosome inherited from one of her parents is coloured red and the chromosome inherited from her other parent is coloured pink. The male common ancestor is blue. The chromosome inherited from one of his parents is coloured dark blue and the chromosome inherited from his other parent is coloured light blue. The second generation are sisters. They share around 50% of their chromosomes IBD. The segments coloured red and pink are segments inherited from their mother and the segments coloured dark blue and light blue are segments inherited from their father. The third generation are first cousins. In each case, the second (white) chromosome derives from their fathers (not shown), the red and pink segments are inherited from their maternal grandmother and the dark blue and light blue segments are inherited from their maternal grandfather. The offspring of these first cousins has segments inherited from both founders on both copies of the chromosome. Where the same segments have been passed down both sides of the pedigree, the offspring of first cousins has extended identical-by-descent tracts or runs of homozygosity.

The effects of inbreeding on health and fitness, termed **inbreeding depression**, have long been noted in plant and animal biology. Wright et al define inbreeding depression as:

> "the detrimental effects of inbreeding, typically causing a reduction in the means of fitness-related traits as a result of increased homozygosity" (Wright, Charlesworth et al. 2003).

Inbreeding is thought to be detrimental to health largely because it increases homozygosity for deleterious **recessive** alleles (i.e. for harmful alleles which are expressed only when homozygous). In terms of complex disease, recessivity is more often partial than complete, that is the effect is expressed when one allele is present, but there is a departure from additivity such that the effect is much greater when two alleles are present (i.e. in the homozygous state).

The probability of homozygosity for a rare allele is:

$q^2(1-F) + qF$ for an inbred organism

q^2 for a non-inbred organism

where:

q is the frequency of the allele in the population

F is the inbreeding coefficient

(Hartl and Clark 1997).

Overdominance, where heterozygotes are fitter than either homozygote (also called **heterozygote advantage** and perhaps best exemplified by Major Histocompatibility Complex alleles and infectious disease) may also contribute to the performance gap between inbred and outbred organisms for some traits, although Charlesworth

suggests that this contribution is relatively minor (Charlesworth and Charlesworth 1999).

1.2 Measuring inbreeding, autozygosity and homozygosity

There are several approaches to quantifying the effects of inbreeding. At a population level, inbreeding can be defined as the proportionate reduction in heterozygosity relative to that expected under random mating (Hartl and Clark 1997). This can be estimated from genotype frequencies:

$$F = (2pq - H_1) / 2pq$$

where:

p and q are the population frequencies of the alleles at a locus

H_1 is the frequency of heterozygous genotypes in the sub-population of inbred individuals.

In terms of assessing the impact of inbreeding on human health, it is important to be able to quantify the reduction in heterozygosity at the individual level. Until recently, this has typically been estimated using pedigree data.

Calculating an individual's inbreeding coefficient from pedigree data involves estimating the probability that both copies of an allele at a randomly chosen locus are autozygous (IBD). There are four steps to this (Hartl and Clark 1997). Firstly, all common ancestors must be identified. Secondly, for each common ancestor, all the genealogical paths leading from one parent up to the common ancestor and back down to the other parent must be traced. These are the paths along which an allele from a common maternal and paternal ancestor could become autozygous. Thirdly, the probability of autozygosity in the individual in question is calculated for each

path in turn. This is straightforward: with Mendelian segregation, the probability that an allele present in a parent is transmitted to a specified offspring is equal to ½. Finally, the inbreeding coefficient is derived by summing the probabilities for each path:

$$F_{ped} = \sum_A (1/2)^i (1+F_A)$$

where

$\sum_A$= summation over all possible paths through all common ancestors

i= the number of individuals in each path

A= the common ancestor in each path

F_A= the inbreeding coefficient of the common ancestor

Since the inbreeding coefficient of the common ancestor is typically unknown, this is assumed to be zero, thus F_{ped} is a relative measure, expressing an individual's degree of inbreeding relative to that of a specified founder generation. This simplifies the calculation:

$$F_{ped} = \sum_A (1/2)^i$$

This approach to estimating inbreeding has clear advantages: where pedigree data are available, it is cheap and easy to calculate; however even where pedigrees are known and accurate, it has two major disadvantages (Carothers, Rudan et al. 2006).

Firstly, meiosis is a random process. Whereas on average, half of the DNA making up a gamete is maternally and half paternally derived, there is a high degree of stochastic variance about this average (Stam 1980; Leutenegger, Prum et al. 2003). As a consequence grandchildren vary in the proportion of DNA they inherit from each of their four grandparents and although mean F_{ped} of the offspring of first

cousins is 0.0625, the standard deviation is 0.0243 (Carothers, Rudan et al. 2006). This variance (i.e. variance as a proportion of F_{ped}) increases with each meiosis (i.e. each degree of cousinship) so it is perfectly possible for the offspring of third cousins to be more autozygous (homozygous by descent) than the offspring of second cousins. Because F_{ped} is derived on the basis of this expectation, it is therefore only a very approximate estimate of individual genome-wide autozygosity.

Secondly, F_{ped} estimates the proportion of an individual's genome that is IBD, relative to that of some poorly characterized founder generation. This generation is usually fairly recent and, moreover, the founders are presumed to be unrelated, when in fact members of historical populations were often related several times over through multiple lines of descent. As a result, this approach fails to capture the effects of distant parental relationships and therefore underestimates autozygosity, particularly in small, isolated populations or in populations with a long tradition of consanguineous marriage (defined as marriage between kin) (Woods, Valente et al. 2004; Liu, Elefante et al. 2006).

With the increasing availability of high-density genome-scan data, interest has grown in exploring whether a valid and accurate estimate of autozygosity might be derived on the basis of genomic marker data. Some of the work in this field has been motivated by those interested in recessive effects in complex disease genetics (Carothers, Rudan et al. 2006); however much of the impetus comes from those searching for specific disease genes using homozygosity mapping. Since the 1980s, many autosomal recessive genes underlying monogenic human diseases have been

identified using homozygosity mapping, which exploits the fact that regions flanking the disease gene will be identical by descent (IBD) in people with the disease whose parents are related to each other (Lander and Botstein 1987). Examples of rare recessive disease genes to have been mapped in this way include Charcot-Marie-Tooth disease (Gschwend, Levran et al. 1996; Saar, Schindler et al. 1998; Rogers, Chandler et al. 2000) and the Fanconi anemias (Bolino, Brancolini et al. 1996; LeGuern, Guilbot et al. 1996; Bouhouche, Benomar et al. 1999; Waisfisz, Saar et al. 1999; Leal, Morera et al. 2001). Botstein and Risch identified nearly 200 studies published between 1995 and 2003 which used homozygosity mapping in consanguineous families to identify rare recessive disease genes (Botstein and Risch 2003). Homozygosity mapping requires an estimate of the proportion of the genome that is autozygous for each affected individual, on the basis of which a LOD score for linkage to a specified locus is computed. Accurate estimation of autozygosity is crucial: under-estimation results in an inflated LOD score and thus false evidence for linkage (Miano, Jacobson et al. 2000; Leutenegger, Labalme et al. 2006). Over-estimation results in false negatives.

Attempts at quantifying individual autozygosity from genetic marker data fall into two categories: single-point and multi-point approaches. The simplest single-point method is termed internal relatedness (IR) (Amos, Wilmer et al. 2001). This is defined as:

$$IR = \frac{(2H - \Sigma f_i)}{(2N - \Sigma f_i)},$$

where H is the number of homozygous loci, N is the total number of loci typed and f*i* is the frequency of the *i*th allele contained in the genotype. This method weights allele sharing by the frequencies of those alleles (Hoffman, Boyd et al. 2004), and thus allows that rare allele homozygotes are given more weight than common allele homozygotes.

The method developed by Purcell et al and implemented in PLINK (F_{plink}) is a variation on IR (Purcell, Neale et al. 2007). For a particular SNP with known allele frequencies p and q, the probability that individual i is homozygous is the probability of being autozygous plus the probability of being homozygous by chance:

$f_i + (1-f_i)(p^2 + q^2)$

If individual i has L_i genotyped autosomal SNPs, O_i is the number of observed homozygotes and E_i is the number of homozygotes expected by chance, then:

$O_i = f_i \times L_i + (1 - f_i) E_i$

which is equivalent to:

$F_i = (O_i - E_i)/(L_i - E_i)$

E_i is estimated using sample allele frequencies, summed across all non-missing SNPs:

$\sum_{j=1}^{Li} 1-2p_jq_j \times T_{Aj}/(T_{Aj} -1)$

where T_{Aj} is twice the number of nonmissing genotypes for snp j.

Related to this is a simple measure of excess homozygosity (H_{ex}), which is equivalent to the numerator of F_{plink}: the difference between the number of observed

homozygotes and the number predicted on the basis of Hardy Weinberg expectation derived from sample allele frequencies.

Carothers et al have proposed another measure of autozygosity, which uses locus-specific heterozygosity to give greater weight to homozygotes at loci with lower expected homozygosity. Expected homozygosity is estimated either from a reference population (which may not be appropriate) or from the actual sample (which may be confounded by unidentified population sub-structure and cryptic kinship) (2006). The key disadvantage to all three of these approaches is that they require the estimation of allele frequencies, a non-trivial problem in many populations (Hoffman, Boyd et al. 2004). A second drawback to these methods is that they assume that individual markers segregate independently.

An improvement in this respect is proposed by Leutenegger et al, whose multi-point approach uses a hidden Markov model to infer autozygosity. Although it is well suited for dense microsatellite maps or mixed microsatellite-SNP maps, it is not in its present form usable with dense SNP maps: it requires that markers are in linkage equilibrium, hence it is computationally complex, requiring either that LD be taken into account or that a subset of SNPs in low LD be selected (Leutenegger, Labalme et al. 2006).

All the above approaches estimate autozygosity by using a variety of inferential methods to filter out homozygous genotypes that have a low probability of being inherited IBD from a recent common ancestor. Methods designed to quantify

homozygosity avoid this complication. These methods were originally developed to estimate heterozygosity. The simplest approach, multilocus heterozygosity (MLH) or observed heterozygosity, simply estimates the proportion of typed loci for which an individual is heterozygous (Charpentier, Setchell et al. 2005). There is, however, a potential for bias if individuals are untyped at particular loci (Coltman, Pilkington et al. 1999). Standardized multilocus heterozygosity (sMLH), defined as the ratio of the heterozygosity of an individual to the mean heterozygosity of those loci at which the individual was typed, avoids this problem (Hoffman, Boyd et al. 2004). MLH is a measure of heterozygosity: the corollary (1-MLH or the proportion of typed markers that are homozygous) is used in this thesis and expressed as H_{pn}.

Any approach to estimating homozygosity or autozygosity on the basis of genomic array data faces the problem that because the markers included in such arrays are highly selected, extrapolation from the homozygosity status of observed markers to the homozygosity status of unobserved markers may not be valid. For example, the estimated probability of homozygosity based on observed common variants cannot be extrapolated to unobserved variants, which may be have very different allele frequencies. This is a particular problem for single-point methods such as H_{pn}, H_{ex} and F_{plink}.

1.3 Inbreeding and health in human populations

The impact of inbreeding on human health is most often explored in the context of consanguineous marriage, which is usually defined as a union between two people related as second cousins or closer ($F_{ped} \geq 0.0156$) (Bittles 2003). Although

consanguineous marriage is now uncommon in most developed countries (fewer than 1% of marriages are estimated to be consanguineous in western Europe, north America, Australasia and Russia), it is still customary in many parts of the world (estimated at between 1% and 10% in the Iberian peninsula, Japan and South America and between 20% and 50% in north and sub-Saharan Africa, and west, central and south Asia) (Bittles 2003). Worldwide, an estimated 20% of the human population live in communities which favour consanguineous marriage and over 8% of children are the product of consanguineous unions (WHO 1985; Bittles 1990).

Whilst many studies investigating the health effects of inbreeding are conducted in populations favouring consanguineous marriage, deliberate consanguinity is not the only mechanism resulting in inbreeding. In small, isolated populations where immigration is negligible, marriage with (distant) kin is unavoidable. The degree of relatedness between two individuals in a population is directly dependent on population size: randomly mating pairs of individuals are inevitably more closely related to each other the smaller the population (Falconer and Mackay 1996). For this reason, small, isolated populations exhibit inflated levels of inbreeding, even where consanguinity is actively avoided. People may consciously choose not to marry a cousin or second cousin; however multiple ancestral relationships between members of the community increases the probability that any marriage within the community is a marriage between (distant) relatives.

It has long been recognised that there is a higher incidence of pre-reproductive monogenic disease, birth defects and early mortality in populations where

consanguinity is common (Khlat and Khoury 1991; Modell and Darr 2002; Bittles 2003). A comprehensive listing of published research is given at www.consang.net (Bittles 2001). More recently, interest has been growing in the impact of inbreeding on the risk of common, complex, late onset diseases with complex genetic and environmental aetiologies. Cardiovascular disease, cancer and adult onset diabetes are the major cause of morbidity and mortality in the developed world and increasingly in the developing world, yet the genetic mechanisms underlying such late onset diseases are still not fully understood. Recent theoretical and experimental advances have lent support to a theory that such diseases result from an accumulation of deleterious mutations of individually small effect throughout the genome. These are kept in check in early life by homeostatic mechanisms and so are not selected against and are allowed to accumulate in the genome. Late onset disease is the result of the combined effects of these alleles and the greater sensitivity of the organism to environmental assaults when compensatory mechanisms break down in later life (Charlesworth and Hughes 1996).

Inbreeding has been associated with increased risk of common late onset diseases (Rudan, Rudan et al. 2003). A number of studies of coronary heart disease (Shami, Qaisar et al. 1991; Puzyrev, Lemza et al. 1992; Ismail, Jafar et al. 2004) and cancer (Simpson, Martin et al. 1981; Lebel and Gallagher 1989; Shami, Qaisar et al. 1991; Rudan 1999) have found evidence suggestive of increased risk associated with close inbreeding. Another approach to the genetic epidemiology of complex disease is to study quantitative physiological traits (QT), the advantage being that this has the potential to provide mechanistic insights into disease pathways. Several studies have

found a small but significant increase in blood pressure (Krieger 1969; Martin, Kurczynski et al. 1973; Hurwich and Nubani 1978; Halberstein 1999; Saleh, Mahfouz et al. 2000; Rudan, Smolej-Narancic et al. 2003; Badaruddoza 2004; Campbell, Carothers et al. 2007) and in LDL cholesterol (Campbell, Carothers et al. 2007) attributable to inbreeding.

The results of such studies should be treated with some caution, as failure to control properly for confounders, particularly socio-economic status (SES) can distort results. Nevertheless, given the global prevalence of consanguinity, investigation of the association between inbreeding and common complex disease is a valid and important subject of epidemiological research in its own right. In addition, it is also important in a more general sense because of what it might reveal about the genetic architecture of common complex diseases which represent the major burden of ill health in the developed, and increasingly in the developing, world.

1.4 The genetic architecture of common complex disease

Over recent years, genome-wide association studies (GWAS) have successfully identified many common genetic variants of modest to large effect size associated with complex disease risk (Florez 2008; Houlston, Webb et al. 2008; McCarthy, Abecasis et al. 2008; Tomlinson, Webb et al. 2008; Frazer, Murray et al. 2009). These findings have been consistent with the "common disease/common variant" or CD/CV hypothesis of disease aetiology, that the genetic component of common disease is the result of genetic variants which occur commonly in the population (Lander 1996). However, despite the increasing success that GWAS have

undoubtedly had in identifying susceptibility loci for a wide range of complex diseases over the last 2 – 3 years, nevertheless, only a small proportion of the heritable component of complex disease and disease-related quantitative traits has been accounted for (Florez 2008; Houlston, Webb et al. 2008; McCarthy, Abecasis et al. 2008; Frazer, Murray et al. 2009). For example, whilst over a dozen susceptibility loci had been published for type 2 diabetes by mid-2008, this accounts for only around 5% of heritability (Florez 2008). One plausible explanation for this might be that many of the alleles contributing to disease susceptibility are of very low frequency in the population. This is consistent with the "common disease/rare variant" hypothesis, which suggests that complex diseases are influenced by large numbers of rare variants distributed throughout the genome (Wright, Charlesworth et al. 2003). The panels of single nucleotide polymorphisms (SNPs) used by GWAS are designed primarily to detect common variants and are not well powered to detect rare variants. For example, it is estimated that the widely used Illumina HumanHap500 panel tags only around 12% of SNPs with minor allele frequency (MAF) of less than 10% (Houlston, Webb et al. 2008; Tomlinson, Webb et al. 2008). A second possible explanation for the limited success to date in accounting for disease heritability might be that disease susceptibility results from the combined contribution of many, many alleles of individually small effect size. Without extremely large sample sizes, current techniques are not able to detect susceptibility alleles with such small effects (Houlston, Webb et al. 2008; Tomlinson, Webb et al. 2008).

The results of GWAS over the last few years suggest that whilst many common disease susceptibility alleles have been identified, this is by no means the whole story. The fact that GWAS have only been able to identify a small proportion of the heritable component of disease aetiology lends weight to the view that rare variants are an important factor in disease susceptibility. The quest for identifying rare susceptibility loci for complex disease and disease traits has led some researchers to perform GWAS in isolate or inbred populations, where random genetic drift can result in high frequencies of alleles that are rare in more cosmopolitan, outbred populations, thus making them easier to identify. A second advantage of isolate or inbred study populations, and of central relevance to this thesis, is that the higher prevalence of inbreeding, and therefore excess homozygosity, in such populations can be used to investigate the involvement of recessive alleles in disease aetiology, leading to a better understanding of the heritable component of disease (Abney, McPeek et al. 2001). The total phenotypic variance (V_t) of a quantitative trait (QT) can be partitioned into environmental and genetic (V_g) components. V_g can be further partitioned into the additive genetic variance (V_a), the component of variance due to genetic effects that are directly transmissible from parent to offspring and which are the main cause of resemblance between relatives; the dominance variance (V_d), the component of variance due to interactions between alleles at the same locus, such as complete or partial dominance or recessivity; and the epistatic variance (V_i), the component of variance due to interactions between alleles at different loci (Wright, Charlesworth et al. 2003). Inbred populations have increased power to detect dominance variance because of increased levels of homozygosity compared with outbred populations (Abney, McPeek et al. 2001). This is essential for the

estimation of broad sense heritability (H^2), defined as the ratio of the total genetic variance of the trait to the total phenotypic variance of the trait (Burton, Tobin et al. 2005). This approach has proven successful: a study in a highly inbred Hutterite population found evidence of dominance variance (i.e. recessivity) in LDL cholesterol and systolic blood pressure (Abney, McPeek et al. 2001; Ober, Abney et al. 2001). This is consistent with the findings of observational studies in inbred populations described above (Krieger 1969; Martin, Kurczynski et al. 1973; Hurwich and Nubani 1978; Saleh, Mahfouz et al. 2000; Rudan, Smolej-Narancic et al. 2003; Badaruddoza 2004; Campbell, Carothers et al. 2007), providing further evidence suggestive of recessive genetic effects on these QT.

1.5 Thesis Aims and Objectives

Recessive genetic effects are usually investigated in the context of inbred or isolate populations. Research has for a long time focused on the health effects of *inbreeding* rather those of *homozygosity*. There are good reasons for this. Firstly, until recently it was simply not feasible to investigate individual homozygosity directly: the only proxy measure available was $F_{ped.}$ Secondly, inbreeding increases homozygosity, making it easier to detect recessive effects in communities with a range of levels of parental relatedness. Studies in inbred communities and kindreds have been extremely fruitful in identifying rare recessive variants causing monogenic diseases (Botstein and Risch 2003). Finally inbreeding results in increased homozygosity for rare recessive alleles, which are predicted to be more likely to be deleterious than common recessive alleles (Wright, Charlesworth et al. 2003).

Recent technological advances have resulted in the development of very dense genotyping platforms, making it possible to develop direct measures of homozygosity and autozygosity. Thus there is now potential for the emphasis of research to move away from inbreeding depression, with all its potentially stigmatising connotations for study populations, to focus directly on investigating the role of recessivity in the genetic architecture of disease.

The broad aim of this thesis is to investigate recessive genetic effects in complex disease aetiology in European populations, focusing not just on inbreeding as a risk factor, but also assessing the contribution to recessive genetic effects of homozygosity which is not attributable to recent parental relatedness.

This study has five objectives:

- To develop a novel genomic measure of inbreeding, which exploits the fact that autozygous genotypes are not evenly distributed throughout the genome but are distributed in runs of homozygosity (ROH) (Figure 1.1).
- To compare this measure with F_{ped} and with other genomic approaches to quantifying both homozygosity and autozygosity.
- To assess the utility of this approach in quantifying the effects on homozygosity levels of deeper demographic history at both the individual and population level.
- To investigate recessive effects, both those attributable to and independent of recent inbreeding, on QT of biomedical importance, using samples from

isolate populations with increased levels of inbreeding compared to more cosmopolitan European populations.

- To investigate recessive effects on colorectal cancer (CRC) risk in non-inbred European populations.

The key study population is the Scottish isolate of Orkney, a remote archipelago off the north coast of Scotland. Genealogical, genetic and physiological data were collected by the Orkney Complex Disease Study (ORCADES), an ongoing, family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in this population. The North Isles of Orkney, the focus of this thesis, consist of a sub-group of ten inhabited islands with census populations varying from ~ 30 to ~ 600 people on each island. Although transport links have steadily improved between the North Isles and the rest of Orkney, the geographical position of these islands, coupled with weather and sea conditions, means that even today they are isolated and that they would have been considerably more so in the past.

Although consanguinity, or marriage between kin, is not the cultural norm in Orkney – indeed there is evidence of consanguinity avoidance during the twentieth century (Brennan and Relethford 1983) – two key factors make the North Isles population ideal for this type of study. Firstly, the North Isles have experienced a period of severe population decline over the last 150 years, fuelled by high emigration and low fertility. The population fell from an estimated peak of 7700 in the 1860s to 2217 by 2001. Secondly, endogamous marriage, or marriage between members of the same

community, was widespread during the nineteenth century and into the twentieth century (Boyce, Holdsworth et al. 1973). Therefore despite consanguinity avoidance, the combined effects of steep population decline and endogamy have led to inflated levels of parental relatedness in the current population.

ORCADES is one of five studies in genetically isolated populations participating in the European Special Populations Research Network (EUROSPAN). The Erasmus Rucphen Family Study (Pardo, MacKay et al. 2005) (ERF; the Netherlands) is a family-based study that investigates the genetic origins of complex diseases in an isolated community in the south west of the Netherlands. The study population essentially consists of one extended family of descendants from 20 related couples who lived in the isolate between 1850 and 1900. The MICROIsolates in South Tyrol study (Pattaro, Marroni et al. 2007) (MICROS; Italy) is a family-based population study that investigates the genetic origins of complex diseases in remote isolate communities in a high valley in South Tyrol. The Northern Sweden Population Health Study (Johansson, Marroni et al. 2009) (NSPHS; Sweden) is a family-based population study which aims to identify genetic and environmental risk factors for common, mainly non-communicable, disease in populations living in the most northerly parish in Sweden. The Croatian study (Campbell, Carothers et al. 2007) (CROAS; Croatia) is a family-based study of residents of small villages in a single Dalmatian island. The village populations of this and neighbouring islands in the eastern Adriatic, Middle Dalmatia, Croatia represent a well characterized meta-population of genetic isolates.

In addition to the five EUROSPAN samples, samples from three other populations are examined in this thesis. The Scottish Colon Cancer Study (Tenesa, Farrington et al. 2008) (SOCCS) is a large case-control study, with cases resident throughout Scotland and controls matched by age, sex and residential postal area. The London colorectal cancer sample is also part of a large case-control study, comprising CRC cases who also have at least one first degree relative affected by CRC and controls with no family history of CRC. Cases and controls are of European origin and from the UK. Finally, the Utah European American sample (CEU) from CEPH (Frazer, Ballinger et al. 2007) consists of 60 unrelated individuals of north and west European origin resident in Utah, USA.

This thesis is divided into seven chapters. Chapter 2 is a detailed description of the ORCADES study population based on genealogical evidence. Data on levels of endogamy and parental relatedness are presented. Chapter 3 develops a novel approach to estimating individual autozygosity from ROH first suggested by Broman and Weber (Broman and Weber 1999) (Figure 1.1). Termed F_{ROH}, this is defined as the proportion of the autosomal genome in ROH above a specified length. Data are presented from a 289,738 SNP panel in 2618 individuals from two isolate populations (ORCADES and CROAS) and two more cosmopolitan populations of European origin (the control sample from SOCCS and the CEU founders from CEPH). Firstly, F_{ROH} is used to compare the four populations, and sub-groups within these populations defined in terms of grandparental endogamy, to see whether this approach can be used to compare populations with different demographic histories: the hypothesis under investigation is that the more isolated the population

or the more endogamous the sub-population, the greater the mean F_{ROH}. Secondly, with the use of high-quality pedigree information available for the ORCADES population, correlations are reported between F_{ROH} and pedigree estimates of autozygosity (F_{ped}). Correlations between F_{ped} and other measures of autozygosity and homozygosity are also shown.

In order to investigate ROH as a disease risk factor, researchers need to be able to detect them reliably. Chapter 4 estimates the shortest ROH that can be reliably detected by SNP panels of different densities and assesses the potential of this approach for quantifying the effects of both inbreeding and more distant individual demographic history. The densest SNP dataset available at the time (HapMap release 23a for CEU founders, containing 3.9 million SNPs, available at www.hapmap.org (HapMap 2002)) was used as a baseline. ROH estimates derived from SNP panels of different densities (500K, 300K and 50K SNPs extracted from the 3.9 million panel to represent the markers available in commonly used off-the-shelf products) were compared with estimates derived from the HapMap release 23a panel.

Using the 5 genetically isolated EUROSPAN populations, Chapter 5 applies F_{ROH} and H_{pn} to assess the role of recessivity in 11 QT of importance in cardiovascular and metabolic disease risk. A linear mixed model, controlled for genetic kinship, was used, to avoid confounding by relatedness between individuals. The effects on phenotypic variance of both inbreeding and homozygosity of more ancient

demographic origin were assessed. A sixth population sample (the controls from SOCCS) was included in the analysis of one QT (height).

Finally, although this analysis shows that ROH longer than 1 to 2 Mb are often indicative of parental relatedness in populations of European origin, shorter ROH (from tens of kb up to 1 or 2 Mb in length) are extremely abundant throughout the genome (Frazer et al, 2007). These result from the inheritance from both parents of identical haplotypes. These may have reached high frequency in the population, perhaps because in the past they conferred some selective advantage or through a process known as allelic surfing, a particular type of genetic drift whereby an allele or haplotype reaches much higher frequency than might otherwise be expected through its presence in the wave front of settlers in a new region (Hofer, Ray et al. 2009). However as haplotype diversity is low in humans, even in the absence of selection or other effects it is not uncommon for identical haplotypes to be inherited from both parents. These shorter ROH have nothing to do with inbreeding as it is commonly understood, although they are strictly speaking autozygous, being inherited from common maternal and paternal ancestors many, many generations in the past. The number, length and location of these shorter ROH differ between individuals and as such, are an aspect of individual genetic variation that might play a causal role in common complex disease and that, therefore, merit further exploration as risk factors in their own right (Lencz, Lambert et al. 2007). Chapter 6 explores this issue by investigating the association between CRC and both F_{ROH} and H_{pn} in the predominantly outbred SOCCS and London CRC case-control study samples, using data from a 500,000 SNP panel.

Chapter 2: Historical and genealogical description of the population of the North Isles of Orkney

2.1 Introduction

A key goal in human genetics and medicine is the development of effective approaches to identifying genetic factors influencing common complex diseases such as heart disease, stroke and diabetes. Diseases such as these result from the cumulative effects of many QT such as blood pressure, blood lipid levels and arterial stiffness. Because QT are less complex than disease itself, they are a more promising target for genetic analysis and can reveal clues about pathways underlying disease process. There are also practical advantages to studying QT: research can be undertaken in general populations rather than being confined to disease cohorts.

The Orkney Complex Disease Study (ORCADES) is an ongoing, family-based study investigating the genetic factors influencing complex diseases through the study of over 100 QT. Orkney is an archipelago of around seventy islands and skerries beginning seven miles north of the northernmost point of the Scottish mainland across the Pentland Firth and extending for about fifty miles northwards (figure 2.1). Seventeen of the islands are now inhabited, a number which has declined over the past 150 years (Collacott 1984). Kirkwall and Stromness, Orkney's only two towns, are situated on the biggest island, known as the Mainland. South of the Mainland are the South Isles of Hoy, Flotta, Burray and South Ronaldsay, the last two connected by road to the Mainland across the Churchill Barriers, built by Italian prisoners of

war in World War II to keep enemy submarines out of the allied harbour at Scapa Flow.

Figure 2.1: Map of Orkney

The North Isles of Orkney, the focus of this study, consist of ten inhabited islands: Westray, Papa Westray, North Ronaldsay, Rousay, Egilsay, Wyre, Stronsay, Sanday, Eday and Shapinsay. Although transport links have steadily improved between the North Isles and Orkney Mainland, with passenger ferry links since the nineteenth century and a regular air service since the 1960s, the geographical position of these islands, coupled with weather and sea conditions, mean that even today they are isolated and not always easily accessible. Strong tidal currents and numerous reefs and shoals hamper easy sea navigation between the isles and prevailing strong winds disrupt both air and sea travel (Collacott 1984).

A number of factors make Orkney an ideal population for localising the genes involved in polygenic disease: Orkney's small population size and isolation mean that there is reduced genetic diversity and reduced heterogeneity for disease-related alleles compared with what would be found in urban populations. Environmental exposures are also more homogeneous in isolate populations, making it easier to identify genetic factors (Wright, Carothers et al. 1999; Ober, Abney et al. 2001). Finally, despite the population decline and increased emigration of the last 150 years, it is still common to find several generations and branches of the same family settled in Orkney. Being able to sample families rather than just individuals means that both linkage and association approaches to variant localisation are possible. The North Isles of Orkney are also well-suited to an investigation of recessive genetic effects on QT. A small population with (until recently) little migration, spread across a number of islands which are (to varying degrees) remote from each other as well as from mainland Scotland, together create the conditions for the inflated levels of

background inbreeding necessary for this type of study. A further advantage for the present study is the availability of reliable and comprehensive records of vital events, making it possible to trace the ancestry of present-day Scots. Civil records of births, marriages and deaths have been universally and systematically produced in Scotland since the mid nineteenth century but even before this time, it was customary for baptisms and marriages to be recorded in parish registers, which are also indexed, digitised and available to researchers. It is therefore possible to reconstruct the pedigrees of present-day study participants back to ancestors living around 200 years ago.

The collection of data from ORCADES study subjects started in 2005, with over 1000 recruits having attended measurement clinics, provided blood for the extraction of DNA and had their pedigrees reconstructed to date. This chapter presents an analysis of inbreeding and endogamy levels in the ORCADES study population derived from pedigree data. The analysis is set in the context of Orkney's population history and in particular in the population history of the North Isles since the eighteenth century. The literature on the genetic origins of modern Orcadians is reviewed, assessing the relative contributions of successive waves of settlers. The availability of historical records of vital events over the last 200 years has given rise to an anthropological approach to population structure, which investigates the influence of major socio-economic forces on marriage, fertility and migration. A review of this literature as it relates to Orkney is presented, providing valuable comparative data to those derived from the present study.

2.2 ORCADES Pedigree Analysis: Methods

The pedigrees of 1071 study participants were traced using official birth, marriage, death and census records held by the General Register Office for Scotland in Edinburgh. Pedigrees were traced back at least 3 ancestral generations (to the level of great grandparents) and up to 8 ancestral generations. Data were entered into RootsMagic, a specialist genealogy programme. Three categories were defined on the basis of grandparental birthplace: those with endogamous ancestry had at least 3 grandparents born on the same isle (isle populations range from ~ 30 to ~ 600); those with mixed Orcadian ancestry had at least 3 grandparents born in Orkney but not on the same isle and those with half Orcadian ancestry had one set of Orcadian-born and one set of Scottish-born grandparents, but with no Orcadian ancestry in the Scottish-born pair. Because people have become increasingly mobile over the last century, grandparental birthplace rather than the birthplace of the subjects themselves was used to create these categories.

Of the 1071 individuals with reconstructed pedigrees, 754 with QT measurements available were chosen for priority genotyping. In order to ensure that the genotyped sample had a wide range of inbreeding levels suitable for subsequent genomic analyses, a mix of individuals was chosen from the three categories described above. In addition preference was given to those with:

- relatives in the study
- full pedigree information to at least 4 or 5 ancestral generations in all lineages (i.e. to the level of great great grandparents or great great great grandparents)
- pedigree evidence of inbreeding

- North Isles ancestry.

Inbreeding coefficients (F_{ped}) were calculated using Wright's Path Method (Wright 1922) – see section 1.2. Mean F_{ped} was estimated separately for the mixed and endogamous categories. Mean F_{ped} was also estimated separately for those with 4 Westray-born grandparents, 4 Stronsay-born grandparents and 4 Orkney-born grandparents, to illustrate the effect of population size on inbreeding.

2.3 Results

725 of the 754 subjects genotyped are included in the following analysis. Apart from failing to meet genotyping quality control (QC) standards (which will be described in the next chapter), subjects were excluded because they did not fit one of the three pedigree categories described above – for example, individuals with one Orcadian and three Scottish parents were excluded, as were those with one set of Orcadian and one set of English grandparents. The mean year of birth of the sample was 1952.

Table 2.1 shows pedigree completion by ancestral generation, where a subject's parents constitute ancestral generation 1, his or her grandparents constitute ancestral generation 2, and so on. Three ancestral generations (i.e. a subject's great grandparents) must be identified in order to detect whether an individual is the offspring of first cousins. Four ancestral generations must be identified in order to detect a second cousin inbreeding loop, and so on. Table 2.1 shows the percentage of subjects in the endogamous and mixed groups with identified inbreeding loops in their pedigrees (the half Orcadian category is not shown, as no inbreeding was

detected in this group). Inbreeding loops originating within 8 ancestral generations were detected in a total of 179 individuals, or in 38.7% of the endogamous group and 9.8% of the mixed group. The proportion of the total sample of 725 individuals with detected inbreeding loops is 24.5%. This is likely to be an under-estimate because of incomplete pedigree data, particularly beyond 5 ancestral generations in the past. The "raw" estimate of mean F_{ped} in the sample, based on all inbreeding loops identified, regardless of pedigree completion, is 0.0019, equivalent to a parental relationship of between third and fourth cousins. This, however, is unlikely to be an accurate estimate, as the increasing proportion of missing ancestral information with each receding ancestral generation means that distant inbreeding loops are likely to be under-estimated. Table 2.1 shows that almost full pedigree information is available to the level of 4 ancestral generations, so estimated numbers of first and second cousin inbreeding loops in the sample are very reliable. Almost ¾ of those in the 5^{th} ancestral generation have been identified; thus estimated numbers of third cousin inbreeding loops are reasonably reliable. Beyond this level, pedigree information becomes increasingly sparse, so the prevalence of more distant inbreeding loops must be extrapolated.

Table 2.1: Pedigree completion by ancestral generation

Number of ancestral generations	% ancestors identified				% of individuals in sample with inbreeding loops detected*			Number of individuals in sample with complete pedigree information
	Endogamous group (n = 390)	Mixed group (n = 286)	Half Orcadian group (n = 49)	Full sample (n = 725)	Endogamous group	Mixed group	Full sample	
3	98.3	98.3	97.4	98.2	0.3	0.7	0.4	655
4	96.8	94.6	89.9	95.5	6.9	3.5	5.1	568
5	83.3	57.6	56.6	71.4	20.8	5.6	13.4	179
6	45.3	27.4	23.9	36.7	31.6	8.7	20.4	7
7	13.7	5.8	5.5	10.0	38.2	9.8	24.4	0
8	1.8	0.8	0.5	1.3	38.7	9.8	24.5	0

** Percentages are percentages of the population where inbreeding loops have been detected originating at or more recently than the ancestral generation indicated.*

Perhaps a better way of estimating F_{ped} would be to use only those subjects with no missing data. Table 2.2 shows mean sample F_{ped} estimates based on 3, 4 and 5 ancestral generation pedigrees using only those with complete pedigree information to these levels. One problem with this approach is that there are several individuals in the sample who are the offspring of first or second cousins but who have otherwise incomplete pedigree information. Excluding these individuals results in an underestimation of mean sample F_{ped}; however including them results in over-estimation. The equivalent F_{ped} values with these individuals included are 0.00029 for 3 ancestral generations (n = 657), equivalent to a parental relationship of 5th cousin; 0.0013 for 4 ancestral generations (n = 584), equivalent to a parental relationship of between 3rd and 4th cousin; and 0.0048 for 5 ancestral generations (n = 247), equivalent to a parental relationship of closer than 3rd cousins. A second problem with this approach is that it cannot estimate inbreeding originating prior to 5 ancestral generations in the past.

Table 2.2: F_{ped} results of subjects with full pedigree information to 3,4 and 5 ancestral generations

No of ancestral generations used in estimate	No of subjects	Mean F_{ped}	Approximate equivalent parental relationship
3	655	0.000095	Between 5th and 6th cousin
4	568	0.00065	Between 4th and 5th cousin
5	179	0.0017	Between 3rd and 4th cousin

Because pedigree information is limited before the early nineteenth century and virtually non-existent before the mid-eighteenth century, perhaps the best way to

estimate the prevalence of inbreeding is to extrapolate on the basis of the number of inbreeding loops and the proportion of ancestors per generation who have been identified. To improve the accuracy of such extrapolations, the sample should first be stratified by population size because simply as a result of population size (assuming high levels of endogamy and low levels of immigration) subjects from larger sub-populations will have lower levels of inbreeding than those from smaller sub-populations. The sample was therefore split into the endogamous and mixed categories, as described above. Those with endogamous ancestry have at least 3 grandparents born in the same isle or parish, whilst those in the mixed category have at least 3 grandparents born in Orkney, but not in the same isle or parish. Table 2.3 shows the number of inbreeding loops detected in each sub-group, and the number expected had all pedigrees been complete. This was estimated by dividing the number of identified inbreeding loops originating from a given ancestral generation by the proportion of ancestors of that generation identified in the sample. For example, if 9 third cousin inbreeding loops were identified but only 30% of ancestors in the fifth ancestral generation had been identified, the estimated true number of inbreeding loops would be 30. Table 2.4 shows raw and estimated true mean F_{ped} statistics for the endogamous and mixed sub-groups and for the sample as a whole. Data are also shown graphically in figure 2.2. In both the sub-groups and the sample as a whole, the raw estimate of inbreeding, which takes no account of pedigree completion, underestimates inbreeding by around 1/3 compared with an estimate based on 8 extrapolated generations. The estimated mean F_{ped} (8 generations) is 0.0029 for the sample as a whole, equivalent to a parental relationship almost as close as 3rd cousins; 0.0041 for the endogamous category, equivalent to a parental

relationship of 3rd cousins; and 0.0016 for the mixed group, equivalent to a parental relationship of a little closer than 4th cousins. F_{ped} in the endogamous group is thus around 2.5 times higher than in the mixed group. Mean F_{ped} estimates on the basis of 3 ancestral generations (i.e. measuring first cousin offspring only) are higher in the mixed than in the endogamous group. This simply reflects the fact that there are two first cousin offspring in the mixed group but only one in the endogamous group (an overall rate of 0.4%). Using data from 4 or more ancestral generations, mean F_{ped} estimates are higher in the endogamous group than in the mixed group, reflecting differences in effective population size between the two groups. Mean F_{ped} estimates increase with each additional generation included in the estimate, reflecting the impact of multiple distant inbreeding loops. This effect is stronger in the endogamous group than in the mixed group because of the impact of the two first cousin offspring on mean F_{ped} in the mixed group. F_{ped} on the basis of 8 ancestral generations is 24 times higher than F_{ped} on the basis of 3 ancestral generations in the endogamous group. In the mixed group, the 8 generation estimate of F_{ped} is 3 times higher than the 3 generation estimate. Although F_{ped} rises with each additional generation of ancestral information, the magnitude of the rise appears to decrease, as the number of additional pedigree loops per generation is counteracted by the fact that they each contribute on average one quarter as much to F_{ped} as loops in the previous generation. This effect is clearer in the endogamous data (figure 2.2) and will eventually lead to an asymptote at the true F_{ped} value.

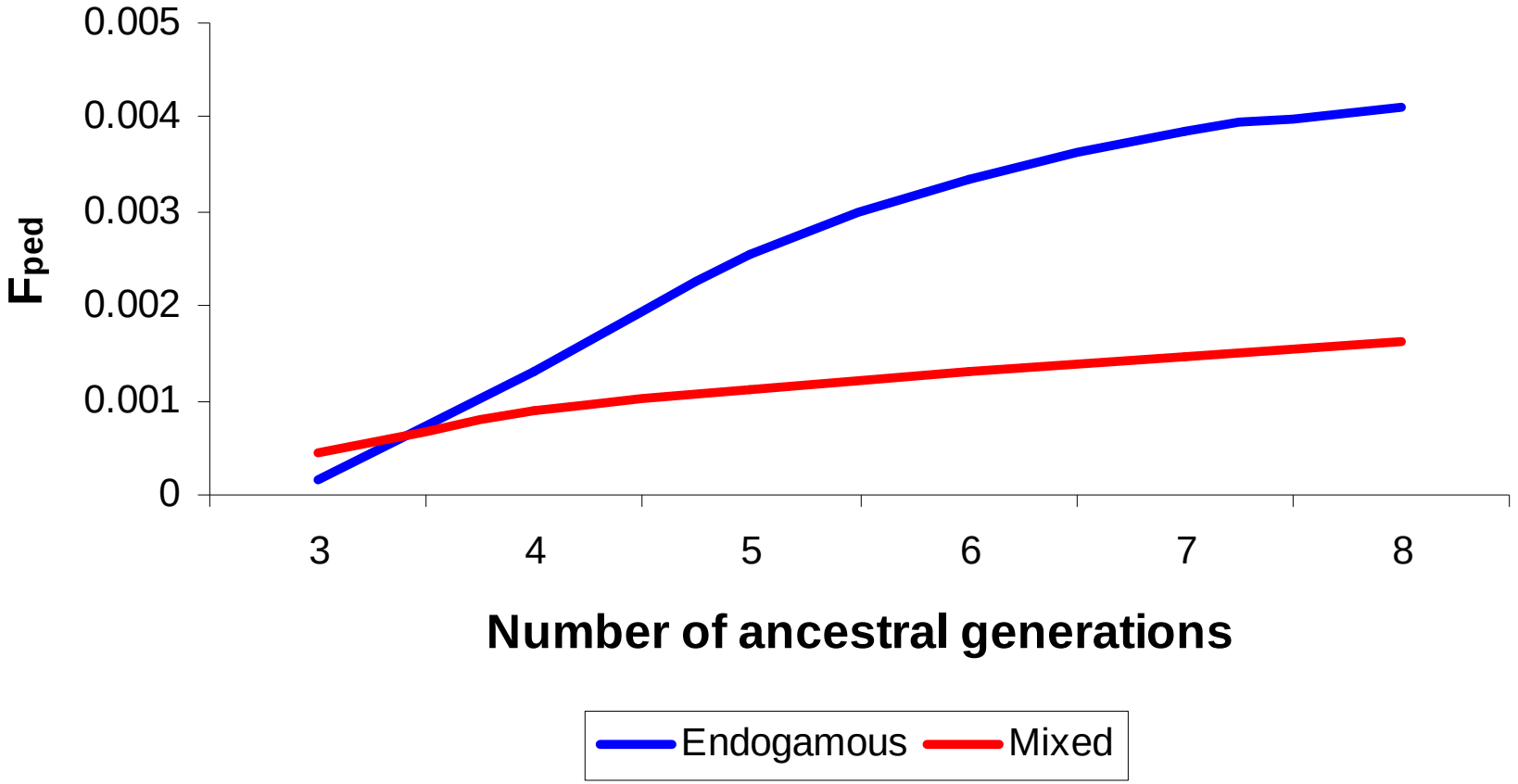
Table 2.3: Observed and expected inbreeding loops by sub-population

Parental cousin relationship	Number of ancestral generations required to detect	Endogamous group		Mixed group	
		Number of loops detected	Estimated true number of loops	Number of loops detected	Estimated true number of loops
1st cousin	3	1	1	2	2
1st cousin once removed	4	4	4		
2nd cousin	4	19.5	20	8	8.5
2nd cousin once removed	5	19	23	3.5	6
3rd cousin	5	66.5	80	2	3.5
2nd cousin twice removed	6	4	9		
3rd cousin once removed	6	44.5	98.5	5	18.5
4th cousin	6	36	79.5	6	22
4th cousin once removed	7	45.5	333	4	69
5th cousin	7	18	132	2	34.5
4th cousin twice removed	8	5	279.5		
5th cousin once removed	8	6	335	2	253
6th cousin	8	1	56	2	253

Table 2.4: Raw and extrapolated true F_{ped} statistics by depth of pedigree and population sub-group

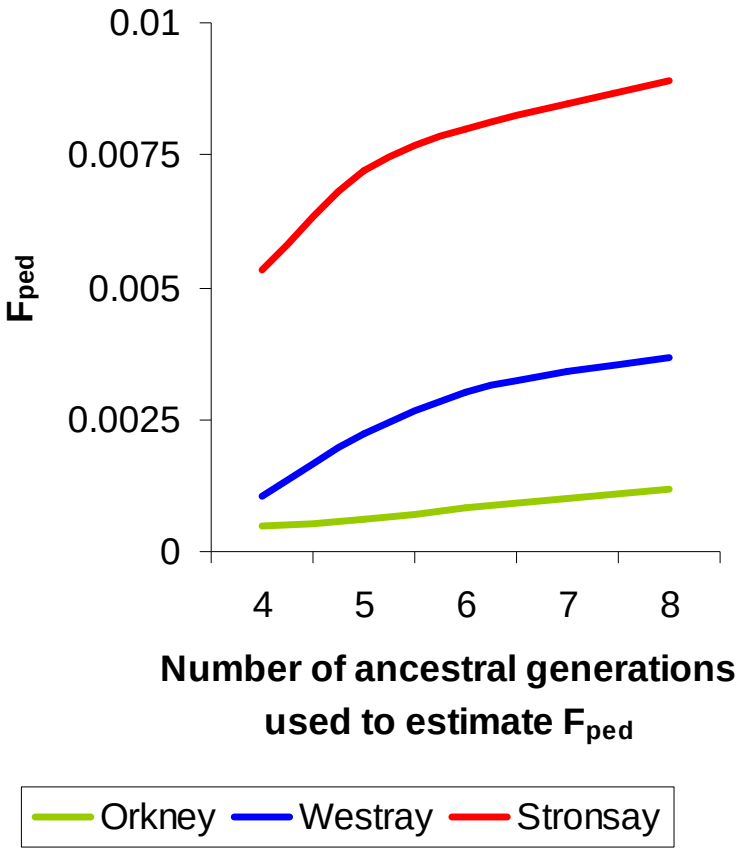
Number of ancestral generations	Endogamous group (n=390)		Mixed group (n=286)		Full sample (n=725)	
	Raw F_{ped}	Estimated true F_{ped}	Raw F_{ped}	Estimated true F_{ped}	Raw F_{ped}	Estimated true F_{ped}
3	0.00016	0.00016	0.00044	0.00044	0.00026	0.00026
4	0.00126	0.00128	0.00087	0.00090	0.00102	0.00105
5	0.00231	0.00254	0.00100	0.00111	0.00164	0.00181
6	0.00266	0.00333	0.00105	0.00131	0.00185	0.00231
7	0.00273	0.00383	0.00106	0.00146	0.00189	0.00264
8	0.00274	0.00411	0.00106	0.00162	0.00189	0.00285

Figure 2.2: F_{ped} estimates by ancestral generation for those of endogamous and mixed Orcadian ancestry



The effect of population size on the prevalence and magnitude of inbreeding can be illustrated by comparing those with 4 grandparents born on the (relatively) large isle of Westray (population today ~ 600, of whom more than 90% are indigenous Orcadian; population in 1841, 1791), those with 4 grandparents born on the smaller isle of Stronsay (population today ~ 300, of whom only ~150 are indigenous Orcadian; population in 1841, 1220) and those with 4 grandparents born in different parts of Orkney (population today ~ 20,000, of whom ~ 70% are indigenous Orcadian; population in 1841, 30,433) (figure 2.3). The mean F_{ped} estimate for those with endogamous Stronsay ancestry is one and a half times higher than the estimate for those with endogamous Westray ancestry, which is in turn more than twice as high as the estimate for those with mixed Orcadian ancestry: in other words, there is an inverse relationship between population size and mean F_{ped}. The 8 generation estimate of F_{ped} for Stronsay is 0.0089, equivalent to a parental relationship of between 2nd and 3rd cousins. The equivalent 8 generation F_{ped} estimate for Westray is 0.0037, equivalent to a parental relationship of 3rd cousins and close to the estimate of the entire endogamous category. The 8 generation F_{ped} estimate for the mixed Orcadian category is 0.0012, equivalent to a parental relationship of 4th cousins. Consistent with figure 2.2, mean F_{ped} in these three groups increases with each additional ancestral generation included in the estimate, although the relative increase in the Stronsay group is less than the relative increase in the Westray and mixed Orcadian groups. The 8 generation mean F_{ped} in the mixed Orcadian group is 1.5 times the 3 generation estimate. The equivalent figure for Westray is 2.5 and for Stronsay, 0.7. The absolute difference between the 3 and 8 generation estimates is highest for Stronsay and lowest for Orkney.

Figure 2.3: F_{ped} estimates by ancestral generations for those with all 4 grandparents born in Orkney, Westray and Stronsay



2.4 Discussion

This analysis provides a clear illustration of the effect of population size on inbreeding. F_{ped} estimates are higher the greater the number of generations included in the estimate, reflecting the fact that first cousin unions are uncommon in this population but more distant cousin relationships, and thus marriages, are more prevalent. Thus inflated levels of background inbreeding are a consequence of small population size and isolation, even in the absence of a cultural preference for consanguineous marriage. This is illustrated graphically in figure 2.4, which shows the estimated reproductive-age population of the isle of Westray (one of the two largest North Isles) since the late 17th century, alongside the number of ancestors of a hypothetical individual born in 1970. Prior to the early eighteenth century, the number of ancestors exceeds the reproductive-age population. Thus, assuming high levels of endogamy and low levels of immigration, inbreeding loops are inevitable in this population (as they are in any population given a sufficiently long view), with the number of loops per individual doubling with each receding ancestral generation. This is a direct consequence of population size, such that individuals with endogamous ancestry from the smaller isles are predicted to have more, and more recent, inbreeding loops than those from Westray. Although the contribution of a single inbreeding loop decreases by ¾ per ancestral generation going back in time (figure 2.5), the existence of multiple loops will inflate inbreeding levels and again, this effect will be more marked the smaller the population.

Figure 2.4: Number of ancestors of an individual born in 1970 compared with the reproductive-age population and total population of the isle of Westray

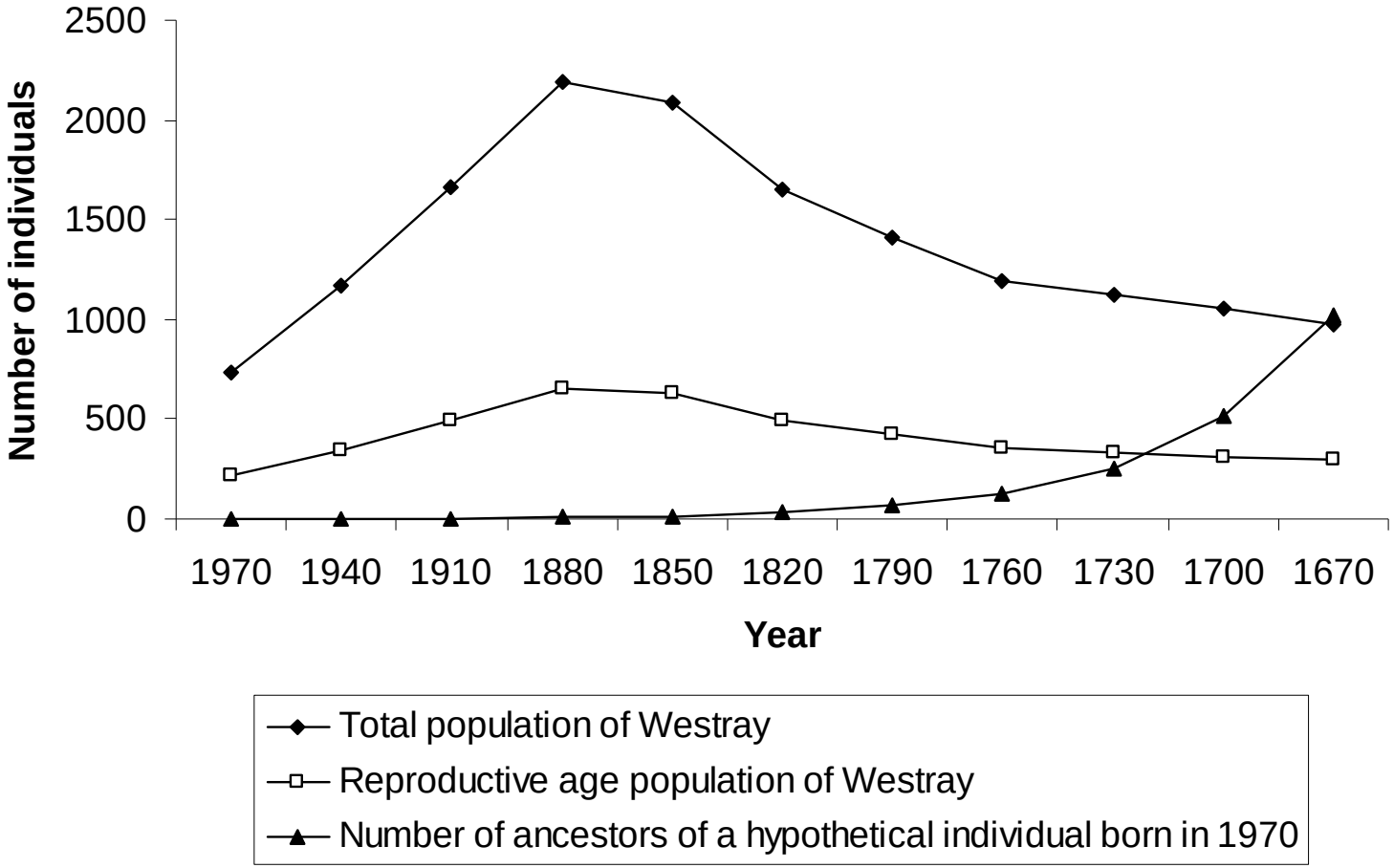
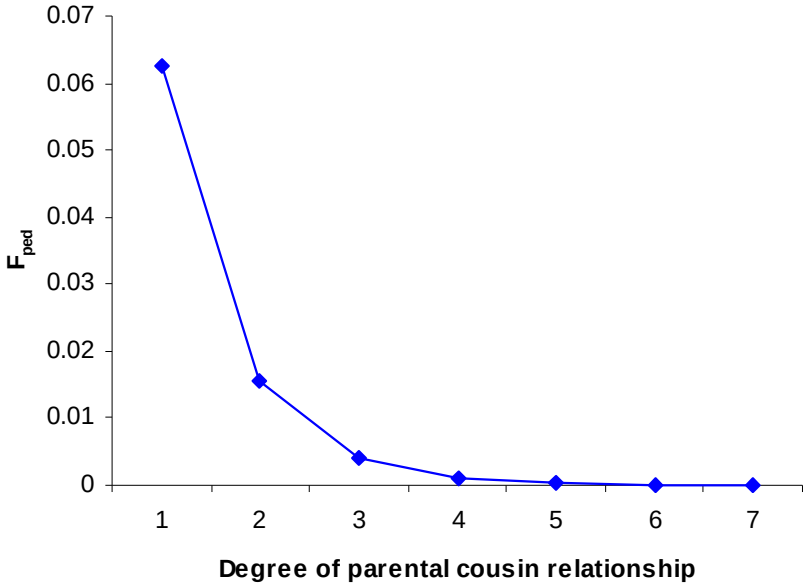


Figure 2.5: F_{ped} of parental cousin relationships



This analysis also illustrates one of the deficiencies of using pedigrees to estimate inbreeding: the reliability of F_{ped} estimates is dependent on the availability of complete pedigree information in all ancestral lineages. Information on births and marriages is relatively complete and reliable for this population; yet even so, it is not possible to trace complete pedigrees beyond five ancestral generations and even this is not possible for every individual in the sample. Whilst extrapolation might provide a reasonable estimate of the mean F_{ped} of the sample going back 8 ancestral generations, this is not possible at the level of the individual. Finally, without good information on historic population size, it is not possible to extrapolate further back than the most distant inbreeding loops identified in the sample. In a small, isolated population, the effect of multiple distant inbreeding loops may be significant, particularly in populations that have experienced a severe bottleneck, for instance through the action of an epidemic.

What can this analysis reveal about levels of inbreeding in today's North Isles of Orkney? The ORCADES study population was chosen to maximise sampling of individuals with at least one grandparent from the North Isles of Orkney. The study population is not, therefore a representative cross-section of the population of Orkney or even of the population of the North Isles. Subjects were not pre-selected, but volunteered to join the study, a factor likely to increase bias. The location of measurement clinics is also likely to have had an effect on recruitment: Clinics were held in four of the North Isles (Westray, Stronsay, Sanday and Shapinsay) and in the town of Kirkwall on Orkney Mainland. This has biased the sample towards those with mixed Orcadian ancestry or endogamous ancestry from Westray, Sanday or

Stronsay. Although individuals with at least 3 grandparents from Birsay, Eday, Egilsay, Fair Isle, Holm, North Ronaldsay, Papa Westray, Rousay, Shapinsay and South Ronaldsay were recruited to the study, numbers were smaller than from Westray, Sanday and Stronsay. Undoubtedly, had it been logistically possible to run clinics on other isles, greater numbers of individuals with endogamous ancestry from these other isles would have been recruited. It appears that the population of Shapinsay has undergone significant immigration from other parts of Orkney, as in spite of recruitment on this isle, only a small number of endogamous volunteers came forward. A probable explanation for this is that Shapinsay is much closer to and more easily accessible from Orkney Mainland compared with the other North Isles, with several scheduled ferries travelling in each direction daily throughout the year.

It is not, therefore, valid to extrapolate from these figures to the F_{ped} of today's Orkney population or even today's North Isles population. Care must also be taken in drawing conclusions about the prevalence of parental relatedness or endogamy on one isle compared with another.

With these caveats in mind, two interesting results emerge. Firstly, given the remote location, low levels of immigration and small population sizes of these islands, it is perhaps surprising that levels of parental relatedness are as low as they appear to be. A possible explanation for this is the tradition of impartible inheritance that prevails in Orkney, whereby only one son inherits the family land and only one daughter receives a dowry. As a result, younger siblings have no choice but to seek their

fortunes elsewhere, a process that has had a dramatic effect on the population of Orkney, as is described in more detail in the following section. The alternative system, whereby the family inheritance is divided between the children, tends to be associated with higher levels of consanguineous marriage, as marrying a cousin is one means of keeping land or property in the family *(personal communication, A.H. Bittles).* A second interesting result to emerge from this analysis is the much higher prevalence of second and third cousin inbreeding loops in the Stronsay group compared with the Westray group. Numbers in the Stronsay group are small and so must be treated with caution, but it is nevertheless plausible to interpret these results in the context of the very sharp decline in Stronsay's population over the past century or so. In the eighteenth century, Stronsay had a thriving kelp industry. As this declined, it was replaced by an equally successful herring fishery. Since this declined from the late nineteenth century, the population has fallen. In the group of 27 individuals with 4 Stronsay-born grandparents, 20 inbreeding loops at the level of 3^{rd} cousin or closer were identified. Such inbreeding, originating from recent generations when the population has been small, appears to account for most of the difference in mean F_{ped} between Westray and Stronsay. Beyond these tentative suggestions, however, it is not possible to draw any broader conclusions about inbreeding or endogamy levels in Orkney. In order to put this analysis in context, it is necessary to look beyond the current study and consider Orkney's population history and the existing literature on endogamy, inbreeding and population structure in these isles.

2.4.1 The population history of Orkney

The earliest evidence of human habitation in Orkney are artefacts left by Neolithic peoples around 4000 BC (Boyce, Holdsworth et al. 1973; Berry 1986). Two distinct, though similar and contemporary cultures have been identified, named after the artefacts they left behind: the Grooved Ware peoples exemplified by the inhabitants of Skara Brae and the Unstan Ware people of Knap of Howar, Papa Westray and elsewhere (Berry 1986).

These ancient people were followed around 700 BC by a sudden appearance in the archaeological record of Iron Age artefacts (Brown 1965; Berry 1986). It is impossible to tell who these Iron Age people were or what happened to the pre-existing population. Modern "indigenist" archaeology tends to the view that the Iron Age people were simply the descendants of earlier inhabitants. This is in contrast to mid-twentieth century "migrationist" archaeology, which saw each new cultural transition as evidence of a new wave of migration. Whatever their origins, these Iron Age people are thought to have evolved into the people described today as the Picts (Berry 1986). It was the Picts who, between 100 and 600 AD built more than 100 brochs, or large fortified stone towers, throughout Orkney. It is thought that each of these brochs housed between 30 and 50 people, suggesting a substantial Iron Age population of 3000-5000 (Boyce, Holdsworth et al. 1973).

During this period, there is also evidence of some Gaelic influence, possibly in the form of Christian missionaries, traders and settlers from Ireland and the Western

Isles and west coast of Scotland. Traces of these people can be found in Orkney place names: "Papa" is thought to be a form of "papae", or priest (Berry 1986).

Probably the single greatest demographic event in Orkney's history, and one which still leaves its mark on place and family names and the local dialect, started around 800 AD with the arrival of the Vikings. It seems that the Norsemen arrived initially in small numbers, perhaps as raiders, but that by the end of the eighth century there was a large movement of population from what is today Norway (Berry 1986). Whether these new settlers annihilated or intermarried with the indigenous population is still a matter of debate. The discovery of Pictish artefacts on Viking archaeological sites (Ritchie 1993), the suggested persistence of Pictish administrative and land tenure systems (Penrith and Penrith 2002) and the persistence of some Celtic place names is cited by some as evidence of assimilation (Penrith and Penrith 2002); however others argue that these could equally plausibly have post-dated as preceded the Norse invasion and that the fact that the overwhelming majority of place and farm names (over 99%) are Norse suggests that the Norsemen overwhelmed and replaced the Picts (Boyce, Holdsworth et al. 1973; Berry 1986; Smith 2001).

By the late ninth century, what had previously been a mere series of settlements emerged as the kingdom of Norway, under a powerful ruling elite. In 875 AD Norway annexed Orkney, establishing it as a Norse earldom, in order, according to the Orkneyinga Saga, to put a stop to raids from exiled Norwegian pirates based there (Collacott 1984). During the following two to three hundred years Orkney was,

while not an independent nation like Iceland, more than just a colony of Norway. Its importance was the result of its strategic position at the heart of lucrative north Atlantic trade routes (Collacott 1984). At the height of Norse power, Earl Thorfinn the Mighty of Orkney held ten Scottish earldoms.

From the thirteenth century onwards, Norse power was waning. During this period there was considerable inter-marriage between prominent Norse and Scottish families, culminating in the first Scottish family taking over the earldom in 1231 (Miller 1986). Orkney's Norse era finally came to an end in 1468, when the islands were pledged to the Scottish crown as part of the dowry of Princess Margaret of Denmark on her marriage to James III of Scotland; Norway and Denmark having been united under the Danish crown in the late fourteenth century. Although the islands were now officially part of Scotland, strong cultural and trading links with Norway persisted. In the sixteenth century, harbour dues were waived for Orkney and Shetland ships berthing in Bergen harbour because they were still considered to be Norwegian and there is documentary evidence of extensive trading between and emigration from Orkney to Norway throughout the seventeenth century (Collacott 1984). The Orkney Norn, the form of old Norse spoken during the Viking period, was gradually replaced by Lowland Scots and largely died out as a distinct language in the eighteenth century, although many words remain in the Orkney dialect spoken today (Flaws and Lamb 1996).

Following the transfer of Orkney to the Scottish crown in 1468, there was an influx of Scottish settlers, with Scottish immigration peaking in the seventeenth century.

There is no reliable evidence as to how many Scots migrated northwards but various attempts have been made to estimate the likely genetic contribution of this wave of settlers by examining surnames. One study of the gravestones of over five thousand individuals identified the three most common surnames in each cemetery. Of these 23 common surnames, 14 were Norse and the remaining 9 were of more recent Scottish origin, implying considerable Scottish genetic intrusion (Boyce, Holdsworth et al. 1973). Another study using the 1614 - 15 Court Book for Orkney and Shetland found that although many Shetlanders had Norse-style patronymic names, in Orkney there were only Scots names or names derived from places. Since it is a Scottish but not Norse custom to name a farmer after his land rather than after his father, this implies considerable Scottish cultural, if not necessarily genetic, influence (Miller 1986). On the other hand, a recent genetic study demonstrated that over half the patrilines of these Orcadian land surnames were Norse in origin (Wilson, Weiss et al. 2001).

There are a number of more recent historical events and processes which have resulted in people moving into Orkney, although not in numbers big enough to make any significant genetic impression. The growth of the herring fishing industry in the nineteenth century saw Scottish fishing boats following the herring on an annual odyssey from the west coast of Scotland, to Shetland, Orkney, Wick and as far south as Lowestoft. The isle of Stronsay became a major herring fishing port, harbouring up to 300 boats during the peak of the nineteenth century herring fishery (Penrith and Penrith 2002). The growth of the Orkney herring industry also prompted migration

from over-crowded Fair Isle in Shetland, with entire families moving south to participate and to teach the Orcadians their skills.

More recently, Orkney was an important naval base in both World War One and World War Two, after which many Orcadian servicemen returned from the south with non-Orcadian wives. The stationing of British servicemen in Orkney during World War Two in particular may have made a genetic impression on the population, although the magnitude of this is unknown (Miller 1986).

The most recent external input to the Orkney population are English families, attracted northwards in steady numbers since the 1960s and 1970s by the promise of a better quality and slower pace of life (Miller 1986). Whether their presence makes a lasting genetic impression will depend on the extent to which the younger generation stay in Orkney and inter-marry with the indigenous population. It will also depend on the relative proportions of outsiders and indigenous Orcadians. Incomers comprise around half the population of Sanday and Stronsay, whilst on Egilsay and Eday, they are so numerous as to outnumber Orcadians.

Finally, two other intriguing, though unverified, sources of genetic input are also worth mentioning. There are persistent rumours in the North Isles of Orkney about sixteenth century connections with the doomed Spanish Armada. After their defeat by the English, the Armada dispersed north into the North Sea, then homewards via the west of Ireland. Many ships were wrecked en route and twenty-seven out of the original one hundred and thirty ships were never accounted for. There are many

Orkney stories about shipwrecked Spanish sailors but perhaps the best known concerns the "Dons of Westray". Legend has it that five or six crew members of a ship wrecked off Fair Isle or, more likely, on the Reef Dyke off North Ronaldsay, ended up in Westray and decided to stay, marrying local girls and taking the local names of Balfour and Hewison, as their own were difficult to pronounce. The Dons allegedly led fairly reclusive lives, not permitting marriage outside the group. Descendants were said to have "Mediterranean features", wavy black hair, short necks and volatile temperaments (Anderson 1988). How much truth there is in this particular legend is unknown, but there were certainly Armada ships wrecked in the waters around Orkney (Anderson 1988). It is difficult but not impossible to distinguish Spanish Y-chromosomes from those of other European origin so it may be possible to identify male patrilineal descendants of Spanish sailors enrolled in ORCADES, if any exist.

A second fascinating but as yet unverified possibility is the genetic contribution of Native American, mainly Cree, women to the Orkney gene pool (Miller 1986). The Hudson's Bay Company was a major employer of Orcadian men in the nineteenth century. At its peak, 80% of those working as traders or explorers for the company in Canada were Orcadian. Men were employed on five year contracts and many never returned, settling in Canada and establishing families with Native American wives. Some, however, did return to Orkney, bringing their wives and families with them (Miller 1986). Cree connections are difficult to pin down from public records - because the native American partner was invariably female, surname analyses are uninformative – however family history information provides evidence of Cree

ancestry in a few families and Native American DNA is highly distinctive. Thus any descendants of Cree women enrolled in ORCADES will be identified when admixture analyses allow recognition of individual Native American haplotype blocks in a European ancestry background. Any matrilineal descendants would be immediately recognisable from their mitochondrial DNA haplotype.

In summary, Orkney's rich history raises many interesting questions. What happened to the pre-Norse inhabitants? Were they annihilated or absorbed by the Vikings? Are today's Orcadians more like modern Scots or modern Norwegians? Is there any evidence for different male and female roles in the various cultural transitions that have taken place in Orkney in the last six thousand years?

2.4.2 The genetic origins of modern Orcadians

These questions have interested biologists and geneticists since the early twentieth century. In 1940, Fisher and Taylor analysed blood group frequencies to test the hypothesis that people from northern Britain were more similar than their southern neighbours to Norwegians because of the history of Viking settlement in northern England and Scotland (Fisher and Taylor 1940). In fact, blood group analysis showed the opposite: moving from south to north within Britain, frequencies of group A decreased, while O increased reciprocally. The southern English sample was closest to that found in modern Norway, whilst the Scottish sample was closest to Icelandic frequencies. Fisher and Taylor concluded that, on the basis of blood group frequencies, modern Norwegians are genetically very different from their Viking forebears because of centuries of infiltration from the east and the north into

Norway. In contrast, they suggest that the Icelandic population, which has undergone little immigration since its establishment by Norse settlers, more closely resembles the original proto-Scandinavian peoples who settled much of the north Atlantic in the Dark Ages.

This early study was obviously constrained by the technological limitations and prevailing genetic theory of the early 1940s. More recent studies have demonstrated clear genetic affinities between Norwegians and Orcadians (Wilson, Weiss et al. 2001). Differences in ABO blood group frequencies are much more likely to be the result of genetic drift or even selection in Norway and/or Scotland. The Icelandic population has also undergone significant genetic drift, such that the postulated close genetic resemblance between modern Icelanders and their Viking forebears is somewhat questionable. Finally, the assumption that those from Scotland and northern England are more Scandinavian than those from southern Britain needs to be challenged: Viking settlement, and thus Scandinavian genetic input, was very localised across Scotland and northern England (Wilson, Weiss et al. 2001).

Until recent technological advancements which have made it possible to analyse genetic origins using DNA markers, studies of genetic affinities had to rely on inferred gene frequency and phenotypic distributions in different populations. Over the last forty years, there have been a number of studies of population distributions of dermatoglyphics (finger and palm prints), taste and colour blindness, pigmentation (Boyce, Holdsworth et al. 1973), blood antigens (Brown 1965; Boyce, Holdsworth et

al. 1973; Welch, Barry et al. 1973; Roberts 1985; Roberts 1986) and longevity (Bowers 1986).

There is considerable historical anecdotal evidence that Orcadians are generally long-lived compared with mainland Scots. There is also evidence that the same is true of Shetlanders and Faroese people (Bowers 1986). Bowers set out to investigate this further by analysing Orkney's civil records for the period 1860-1964. She found no evidence of exceptional levels of extreme longevity but she did find evidence of a high modal age at death for both men and women: throughout the period, the modal age of death for both sexes was in the 75-84 age band, significantly higher than that in comparative British populations. Bowers argues that there is strong evidence for a genetic rather than environmental basis for this: the pattern is consistent across parishes and throughout the period, despite changes in medical care, health and hygiene practices and standards of living which have dramatically increased mean life expectancy (Bowers 1986).

Studies of physical characteristics have produced little beyond concluding that Orcadians are somewhat more blue eyed than their mainland neighbours but not as blonde as Scandinavians (Boyce, Holdsworth et al. 1973). Studies of blood group gene and phenotypic frequencies have fared a little better, but have been hampered by inconsistent comparative data, the limited variation between Scottish and Scandinavian gene frequencies and conflicting evidence from different markers (Welch, Barry et al. 1973). It is also important to note that these are not simply markers of genetic affinities: the effect of selection on blood antigens should not be

discounted. In summary, these studies found evidence of lower frequencies of blood group A in Orkney than in Norway and Denmark, although the frequency in Orkney was higher than in the rest of Scotland (Brown 1965; Boyce, Holdsworth et al. 1973). There were also consistent findings that a high frequency of blood group B distinguished Orkney both from more southerly parts of Britain and from other Norse areas (Brown 1965; Boyce, Holdsworth et al. 1973). Brown also found affinities between Orkney and Scottish east coast fishing communities, which share unusually high frequencies of group B (Brown 1965). In a multivariate analysis of a range of blood gene frequencies, Roberts concluded that Orcadians are distinct from the Gaelic Atlantic populations of northwest Europe; are genetically closer to North Sea than Atlantic populations and have genetic affinities with other Viking areas of Britain, such as Newcastle and Cumbria (Roberts 1986).

Various studies have found extreme values for some markers (Welch, Barry et al. 1973; Roberts 1986). Roberts suggests that such outliers might represent traces of pre-Norse inhabitants. More likely, unusually high allele frequencies may be of more recent origin, resulting from the effects of genetic drift in these small, isolated, island populations. Most notably, a high frequency of a rare variant of superoxidase dismutase has been found in Westray (Welch and Mears 1972) but not in nearby North Ronaldsay (Welch 1973). From worldwide distributions of this variant, it has been hypothesised that it is of Scandinavian origin and spread through Viking migrations (DeCroo, Kamboh et al. 1988).

Recent technological developments have meant that much more sophisticated population genetic analyses can now be brought to bear on the questions of Orkney population affinities. Very high resolution systems which can identify many different genetic types and therefore allow finer scale inference are now available. In particular, non-recombining Y-chromosome and mitochondrial DNA markers, which can be used to infer population history even in the face of admixture, have been developed. These can be used to distinguish different population genealogies and to explore differences between male and female roles during periods of major cultural transition. Wilson and colleagues used this approach to place modern Orcadians on a genetic map of Europe and to explore the extent to which the Norse invasion involved male and female population movement from Scandinavia (Wilson, Weiss et al. 2001).

Comparing Orcadian Y-chromosomes with those from modern Norway (representing the Norse source population), Anglesey and Ireland (Celtic/pre-Anglo-Saxon British), West Friesland (Anglo-Saxon) and the Basque region of Spain (pre-Neolithic European), they found that, unsurprisingly, Orkney was situated mid-way between the Norwegian and Celtic samples. Because surnames co-segregate with the Y-chromosome, the Orkney sample was further subdivided according to whether subjects' surnames were of Norse or Scottish origin. Whereas the Y-chromosomes of those with Scottish surnames were indistinguishable from either the Welsh or Irish samples, 38% of those with Norse surnames had Y-chromosomes of unambiguously Scandinavian origin, providing genetic confirmation of the movement of males from Scandinavia during the Norse period. The more ambiguous Y-chromosome

haplotypes are those that are found commonly in both Scotland and Norway, thus it is impossible to say for certain with the given level of resolution whether or not an individual Orcadian man with such a haplotype is patrilineally descended from a Viking. It is likely, however, that a large proportion of these ambiguous haplotypes are also of Norwegian origin. This is because people do not migrate according to their haplotype, so ambiguous and unambiguous Norwegian haplotypes should have been brought to Orkney by the Vikings in proportion to their frequency in the source population. Newly discovered markers which divide the ambiguous origin groups into subgroups with more circumscribed distributions will soon allow much improved assignment of patrilineal ancestry in Orkney.

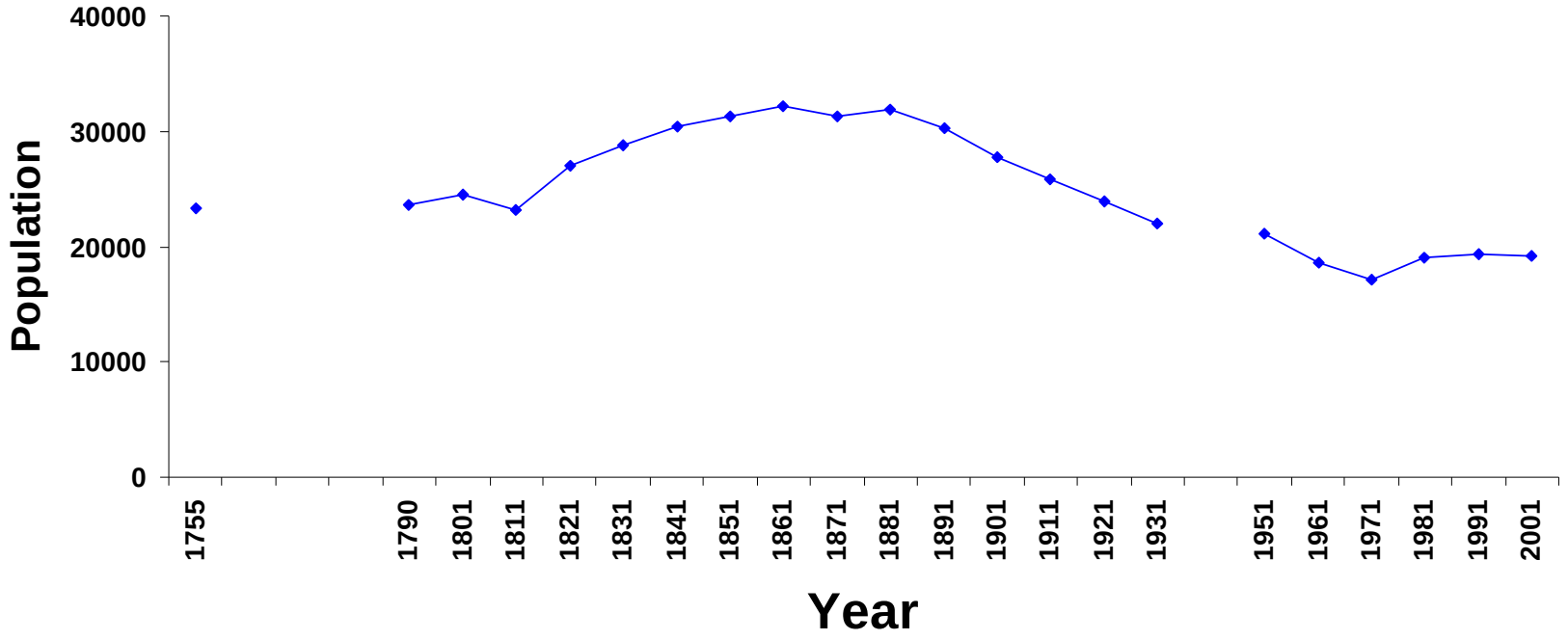
Wilson and colleagues suggest that it is not possible to assess the relative contributions of Celtic and Norse female ancestry to the modern Orcadian population. This is because the homogeneity of mitochondrial DNA from European source populations (probably as a result of female migration during at least one earlier major cultural transition in Europe) means that there is no power to distinguish Norse and Celtic strands at the level of resolution utilised in their paper (Wilson, Weiss et al. 2001). Whole mtDNA molecule sequencing should provide considerably more information but at present, the extent to which the Norse invasion was an exclusively male affair and the question of whether the Vikings annihilated or fused with the indigenous population, remain unresolved. Another study estimates the mitochondrial DNA make-up of modern Orcadians as 35.5% Scandinavian (95% confidence interval 13.0 – 64.5) and 64.5% Celtic (95% confidence interval 35.5 – 87.0) (Helgason, Hickey et al. 2001). This is consistent with the involvement of

female Norse settlers, but is uninformative about the identity or fate of the pre-Norse population.

2.4.3 Orkney since the eighteenth century

There are few reliable data on the population of Orkney prior to the mid-eighteenth century (Collacott 1984). For the first hundred or so years for which data are available, numbers grew steadily, peaking at over 32,000 in the 1860's. There then followed a period of sharp decline, to a low point of around 17,000 in the early 1970s. Since then, the population has recovered slightly as a result of English immigration and greater economic prospects servicing the North Sea oil industry (Miller 1986) (figure 2.6).

Figure 2.6: Population of Orkney: 1755-1971



Source: UK Census

Although civil registration of births, marriages and deaths has only been compulsory since 1855, most parishes kept fairly comprehensive records of births/christenings and marriages from at least the mid-eighteenth century onwards. Thus Orcadians can generally trace most of their ancestors back to those born between 1750 and 1800. The availability of such detailed records makes it possible to examine population trends at the micro level. Individual decisions about whether, whom and when to marry, how many children to have and whether to migrate have a direct bearing on population size and structure. These decisions are in turn influenced by prevailing socio-economic conditions, customs and social attitudes, religion and, of course, geography. This section describes key historical events and processes which have had a bearing on population structure over the last 250 years. It then outlines the findings of a number of studies which have investigated the effects on population size and structure of geographic isolation and marriage and kinship trends, with particular reference to consanguinity and endogamy.

Population growth and decline

From the 1750s to the 1860s the population of Orkney grew steadily. Rich soil and the introduction of agricultural improvements by major landowners meant that the islands were able to sustain this growing population, whilst avoiding much of the upheaval experienced by crofting communities elsewhere in the Highlands and Western Isles at this time (Collacott 1984). The introduction of commercial steam shipping in the 1830s, both inter-island and between Orkney and mainland Scotland, meant that farmers could switch from producing uneconomic and risky grain to much more lucrative cattle, in the knowledge that there was now a reliable means of export

(Collacott 1984). Orcadians also responded successfully to the Industrial Revolution by turning whole communities over to the production of kelp from seaweed, which was used to make potash for Glasgow's soap and glass industries (Penrith and Penrith 2002). At its peak in the late eighteenth century, the kelp industry had overtaken agriculture in economic importance in some parts of Orkney (Collacott 1984). After the kelp industry collapsed in 1820, effort was diverted into developing a commercial herring fishing industry, which was important to the islands' economy into the early twentieth century (Collacott 1984).

After peaking at over 32,000 in the 1860s, Orkney's population, in common with rural populations throughout western Europe (Brennan and Relethford 1983) went into a period of steep and sustained decline, reaching a low of around 17,000 in the 1970s, a process characterised by falling birth rates and rising emigration. Whilst no single factor was responsible for this trend, a number of inter-related causes can be identified. The enclosure and division of common grazing land made crofting a less viable way of life and rising living standards increased the minimum farm size regarded as sufficient to sustain an acceptable standard of living (Collacott 1984). The tradition of impartible inheritance, whereby only one male heir inherited land and usually one daughter received a dowry, meant that younger siblings had either to emigrate or seek work as landless labourers (Brennan and Dyke 1980). Agricultural mechanisation meant fewer employment opportunities for such people (Brennan 1981), whilst universal education and improved communications links with the outside world brought them the possibility of new opportunities elsewhere. Whilst young men tended to emigrate abroad, women typically moved to the UK mainland

to work as domestic servants (Coull 1966). With the waning of the herring industry, Orkney seafarers also took jobs further afield, with the Greenland whaling fleet and the Royal Navy (Penrith and Penrith 2002). These economic changes were accompanied by a process of social modernisation. Whereas in earlier times, marriage had been primarily an economic contract between two families, it was now increasingly regarded as a decision between two individuals (Brennan 1981).

Population trends in the North Isles of Orkney

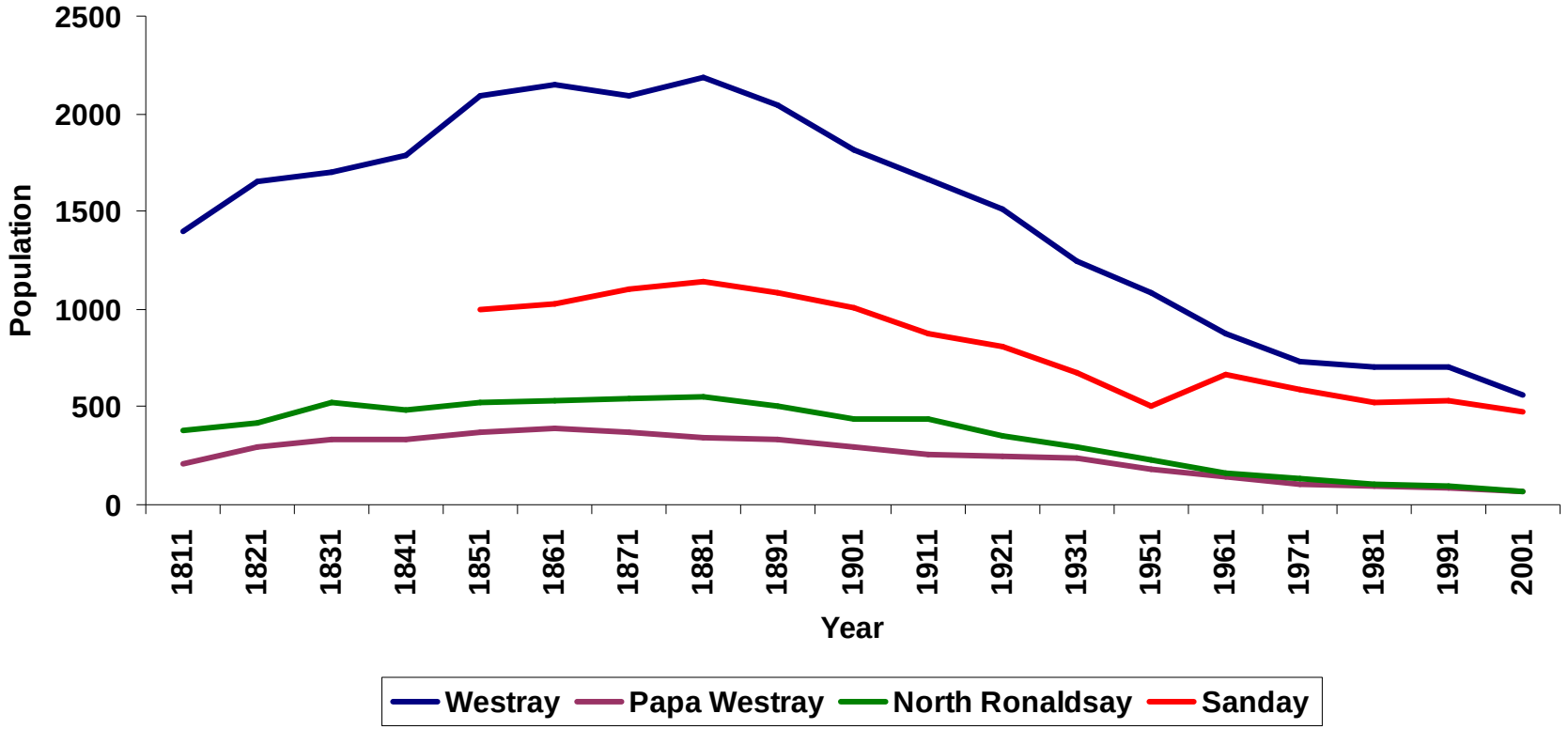
How did these trends affect the North Isles of Orkney? Figure 2.7 shows the populations of Westray, Papa Westray and North Ronaldsay from 1811 and the population of Sanday from 1851. In general, the pattern in the North Isles follows that in Orkney as a whole, but with populations peaking slightly later, followed by a considerably more extreme decline, as a result of the combined effects of decreasing fertility and increasing emigration and celibacy (Brennan 1981). In fact, depopulation resulted in the desertion of several of the North Isles: Linga Holm and Eynhallow have been uninhabited since 1851 and Faray since 1931. Gairsay and Papa Stronsay were deserted in 1951 and Auskerry in 1961 (Collacott 1984) although these last three have since been re-inhabited, albeit by non-indigenous Orcadians.

In Orkney as a whole, the only places to buck this trend up until the 1960s were the towns of Kirkwall and Stromness, which actually grew in size between 1861 and 1961 at the expense of the smaller isles (see figure 2.8) (Boyce, Holdsworth et al. 1973). In a study of population trends on Westray, Coull observed a temporal

change in migration patterns: whereas the majority of pre-World War Two Westray emigrants headed abroad, after World War Two most emigrated to Orkney Mainland, either to work as farm labourers or in Kirkwall's burgeoning twentieth century service sector (Coull 1966).

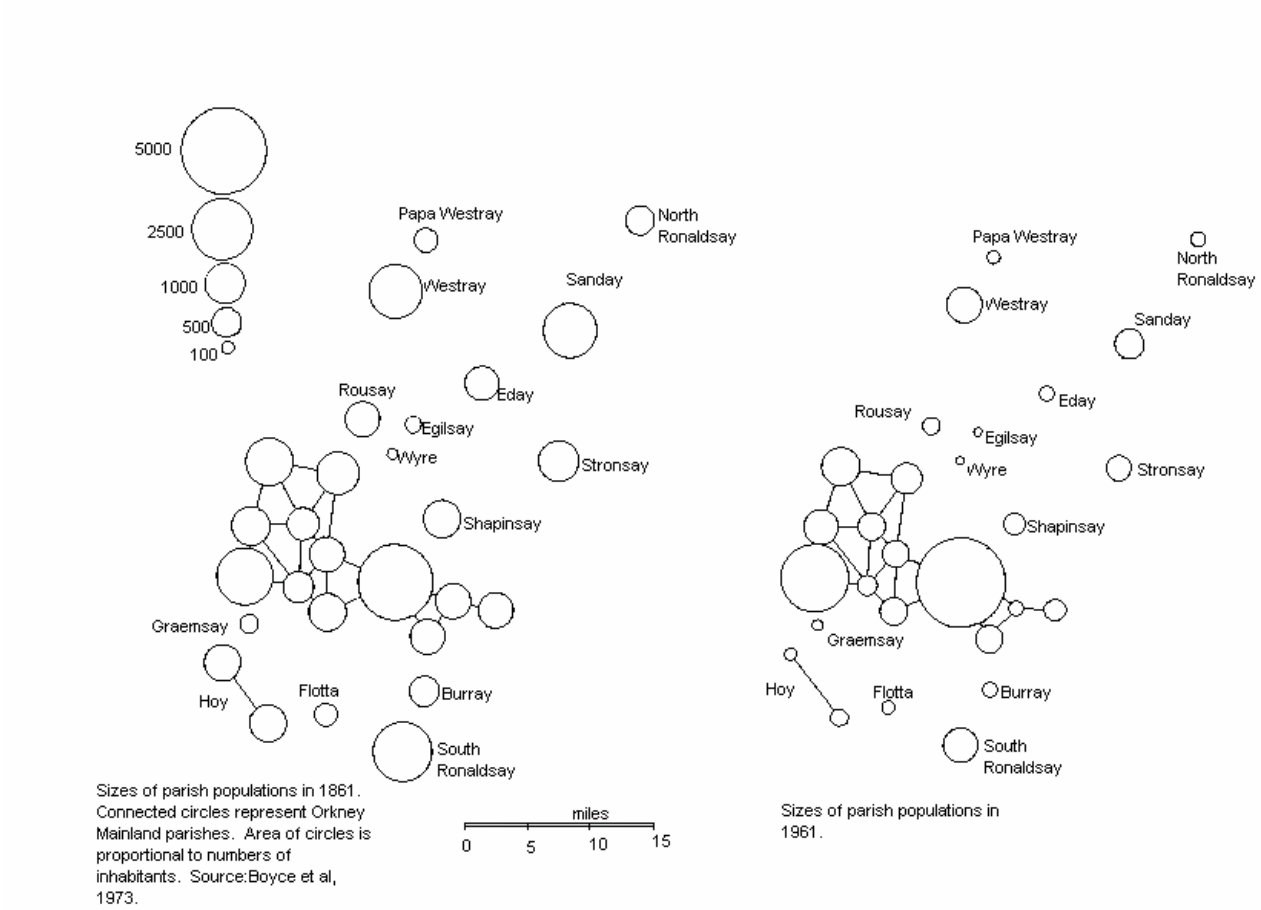
A second interesting pattern to develop over this period involved changes in relationships between the North Isles. In a study of marriage patterns on Sanday between 1800 and 1964, Brennan noted a temporal increase in the proportion of spouses born off-island, coupled with an increase in off-island spouses coming from the larger population centres of Kirkwall and Edinburgh, as opposed to the other North Isles. Thus she suggests that over this period the North Isles were becoming more isolated from each other (Brennan 1981). Similar trends were observed in Westray (Collacott 1984).

Figure 2.7: Populations of Westray, Papa Westray, North Ronaldsay and Sanday, 1811 - 2001



Source: Census returns

Figure 2.8: Change in population size of Orkney parishes, 1861-1961



Sizes of parish populations in 1861. Connected circles represent Orkney Mainland parishes. Area of circles is proportional to numbers of inhabitants. Source:Boyce et al, 1973.

Sizes of parish populations in 1961.

Effects on population structure

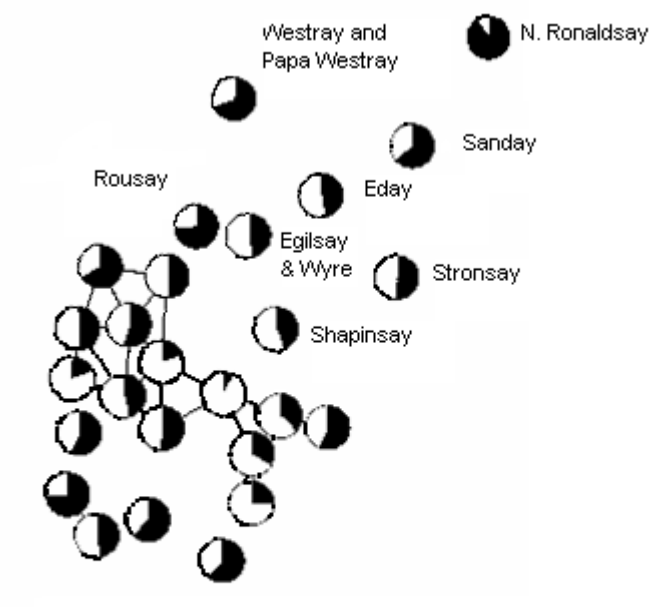
How did the combined effects of sharp population decline and increasing mobility impact on marriage behaviour in the North Isles and how has this in turn affected population structure? These issues have been extensively investigated by Brennan and colleagues on the isle of Sanday (Brennan 1981; Brennan, Leslie et al. 1982; Relethford and Brennan 1982; Brennan and Relethford 1983). Other studies have reported on endogamy in the North Isles (Boyce, Holdsworth et al. 1973; Collacott 1984), on levels of inbreeding in Orkney (Roberts, Roberts et al. 1979), on population substructure (Roberts 1985) and on demographic factors impacting on population structure (Brennan and Dyke 1980).

Endogamy

Endogamy is the social practice of marrying within the same clan or community. For the present purposes, it is defined as marrying within the same isle or parish. A detailed investigation of marriages in Westray between 1855 and 1974 found that in 87% of these, both partners lived in Westray before marriage. Sixty-three per cent of Westray marriages were between people who lived less than 3 miles apart (Collacott 1984). Eighty-four point five per cent of marriages in Sanday between 1855 and 1965 were between Sanday residents and 57.8% were between people living within three miles of each other (Boyce, Holdsworth et al. 1973). It should be noted, however, that residence before marriage is a less reliable measure of endogamy than natal residence because during this period it was common for young single people of both sexes to move away from home in search of employment (Coull 1966). Figure 2.9 is an illustration of endogamy levels in different parts of Orkney in 1861. It

shows proportions of marriages in which both partners were born in the same parish (Boyce, Holdsworth et al. 1973). By this measure the outer isles were generally more endogamous than Orkney Mainland (connected circles), although the more remote westernmost (Birsay) and Easternmost (Deerness) parishes had similar endogamy levels to the North Isles. Using natal residence on Sanday as a measure of endogamy, Relethford and Brennan found lower levels than those quoted for Westray and Sanday above. Nevertheless, 65% of marriages in the mid-to-late nineteenth century were between Sanday-born couples. Although this fell steadily over the next century, a majority (52%) of mid-twentieth century Sanday marriages were still endogamous (Relethford and Brennan 1982).

Figure 2.9: Proportions in 1861 of marriages in which both partners were born in the same parish



Proportions (shaded areas of circles) in 1861 of marriages in which both partners were born in the same parish. Connected circles are Mainland Orkney parishes. Source: Boyce et al, 1973.

Consanguinity

Four key questions about consanguinity in Orkney since the late eighteenth century are:

- Can any temporal trends be observed?
- Is there any evidence of consanguinity avoidance or preference?
- What is the association between consanguinity levels and population size?
- Has religion influenced attitudes to consanguineous marriage?

The coefficient of kinship (Φ) is defined as the probability that two random alleles from the same locus in two individuals are inherited IBD from a common ancestor (Relethford and Brennan 1982). An individual's inbreeding coefficient (F_{ped}) is the equivalent of the kinship coefficient between his or her parents. Defining consanguineous marriage as marriage between second cousins or closer, Brennan found that the kinship coefficient of married residents of Sanday increased from an average of 0.001481 in the period 1800-1854, to 0.001962 in the period 1885-1924. It then declined to zero in the period 1925-64 (Brennan 1981). This study also found that whilst absolute numbers of marriages between relatives remained fairly constant between 1800 and 1924, the proportion involving close relatives fell and that involving more distant relatives increased reciprocally. Interestingly, although this study looked at marriages from over 150 years ago, kinship levels are similar to the mean F_{ped} of the ORCADES endogamous sub-group (0.0013 based, like the Brennan sample, on 4 ancestral generations).

Roberts and colleagues conducted an Orkney-wide investigation of inbreeding levels, covering individuals born in the period 1870 – 1949. This was based on complete 6 -

8 generation pedigrees. Mean F_{ped} was 0.001834, comparable to the raw F_{ped} of the ORCADES sample based on 7 ancestral generations (0.0019), but considerably lower than the estimated true 7 generation ORCADES estimate of 0.0029. Roberts found evidence of some level of inbreeding in 7% of the sample, considerably lower than the 24.5% of the ORCADES sample with detected inbreeding loops. Unlike Brennan, Roberts found no evidence of change over time but because the time periods of the two studies are different it is difficult to draw any conclusions from this. Roberts and colleagues also investigated regional differences in the prevalence and levels of inbreeding within Orkney, although sample sizes for this analysis were small. The most inbred areas were the West Mainland parishes (12.1% inbred; mean F_{ped} = 0.00325) and the outer North Isles (Westray, Papa Westray, Eday, Sanday, North Ronaldsay and Stronsay – 8.2% inbred; mean F_{ped} = 0.003189). The inner North Isles of Rousay, Egilsay, Wyre and Shapinsay had an inbreeding prevalence of 6.7%, mean F_{ped} = 0.001042 and the South Isles had a prevalence of 7.9%, mean F_{ped} = 0.000848. No inbreeding was detected in the East Mainland, where Kirkwall is located, and this was significant (p = 0.01) (Roberts, Roberts et al. 1979).

A measure of attitudes towards consanguinity can be derived by comparing the average kinship between actual spouses with that between maters and their potential mates (Brennan and Relethford 1983). This analysis has to be controlled for the effects of geography: because marriage was typically conducted across very short distances (i.e. people tended to marry their neighbours) and because levels of kinship declined with distance (i.e. relatives tended to live close to each other) it is important to distinguish between a tendency to marry one's neighbours and a preference for kin

marriage. Brennan and Relethford analysed Sanday marriages in three time periods (1855-84, 1885-1924 and 1925-64) in order to elucidate temporal trends in relatedness and marital distance. They found that in the two earlier periods, Sanday-born males preferentially married their relatives living within 10km. In other words, they married relatives more than would be expected, taking the pool of potential mates as those living within 10 km. In the earliest period, they found evidence for consanguinity preference even for relatives born outside Sanday. In contrast, from 1925 onwards, they found evidence of consanguinity avoidance at all marital distances (Brennan and Relethford 1983).

In many European countries, prevailing religious influences have had a profound effect on the prevalence of consanguineous marriage. For Roman Catholics, Diocesan dispensation is still required for first cousin marriages and prior to 1917, this was also required for second and third cousin marriages (Bittles 2001; Bittles 2003). In contrast, there is no such prohibition for members of Protestant denominations and consequently, the prevalence of consanguineous marriage has been higher in predominantly Protestant European countries than in countries where Roman Catholicism is the dominant religion. As a predominantly Protestant region, there has been no religious prohibition against kin marriage in Orkney.

In order to investigate the inter-play of mating decisions and population size on genetic variability, Brennan and colleagues analysed population data from Sanday using a model which quantifies the genetic effects over time of these factors. They suggest that whilst consanguinity avoidance through choosing unrelated spouses will

have a modest effect on maintaining genetic variability over time, this is insignificant compared with the opposite effect of high levels of emigration and celibacy, which act to reduce the effective population size, thereby increasing random inbreeding and dramatically reducing genetic variability in the long term (Brennan, Leslie et al. 1982).

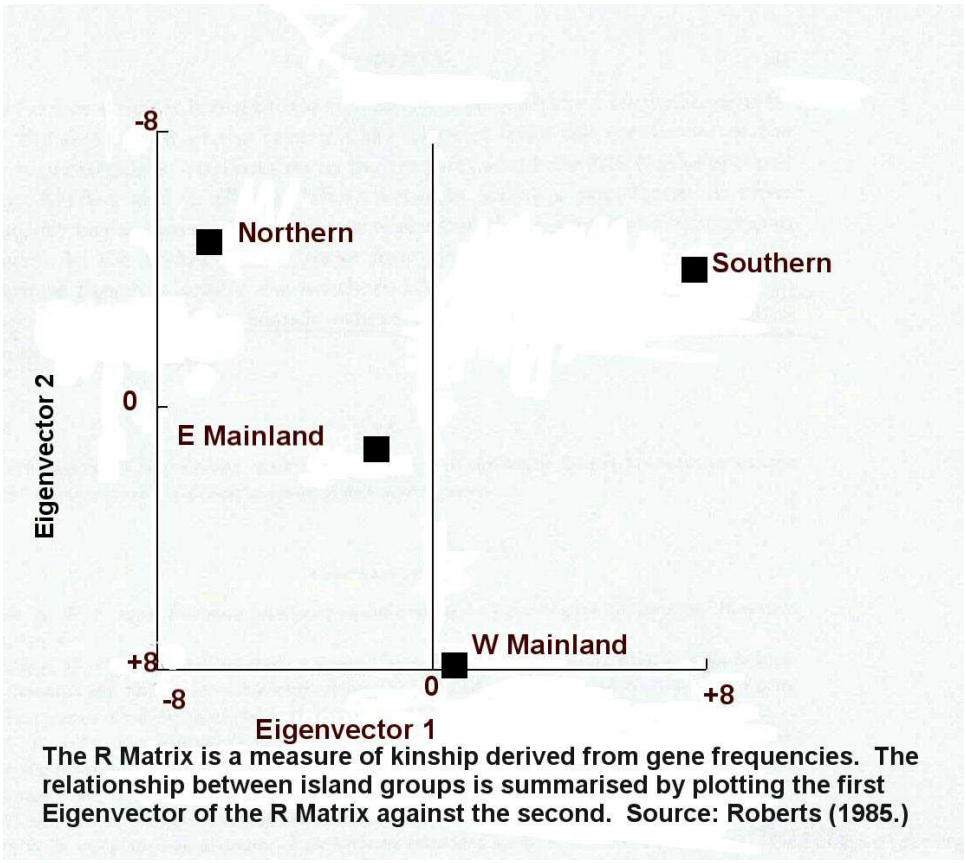
Of course, real life is more complex than population models and empirical studies have demonstrated a decrease in inbreeding levels associated with population decline (Brennan and Relethford 1983). This is because increased prosperity and ease of transport, coupled with changes in social attitudes towards marriage with kin, have enabled people to look further afield for marriage partners, as evidenced by the increasing proportion of immigrants amongst married couples in Sanday (Brennan 1981).

Evidence of population sub-structure

High levels of endogamy and the geographic isolation of Orkney's outlying islands prompted Roberts to analyse genotype frequencies for red cell blood groups, red cell isoenzymes, serum proteins and HLA types in order to investigate the extent to which there was detectable genetic structure between four island groupings: the North Isles, the South Isles, West Mainland and East Mainland (Roberts 1985). If there were no population substructure, individuals would be equally likely to choose a partner from their own island group as from any other island group and genotype frequencies across the entire sample would be expected to be in Hardy-Weinberg Equilibrium. The extent to which actual genotype frequencies diverge from Hardy-

Weinberg Equilibrium is a measure of population substructure. Roberts found evidence of significant excess homozygosity, suggesting that there is indeed population structure within Orkney (see figure 2.10). A multivariate principal components analysis of a number of markers using different measures of genetic distance suggested that the greatest genetic distances appeared to be between the northern and southern isles and between the North Isles and the west mainland. Furthermore, genetic distances within Orkney were greater than between Orkney and mainland Britain. Finally, Roberts used gene frequency data to estimate levels of inbreeding in the sample. At 0.00771, the mean level of inbreeding calculated from gene frequency data was over four times that estimated from a 200 year old pedigree in the earlier study quoted above (Roberts 1985). This suggests that the population substructure revealed here dates back considerably further than 200 years. This study also provides further evidence of the inadequacy of F_{ped} at capturing the effects of multiple distant inbreeding loops in a small population.

Figure 2.10: Fine-scale population structure in Orkney



The R Matrix is a measure of kinship derived from gene frequencies. The relationship between island groups is summarised by plotting the first Eigenvector of the R Matrix against the second. Source: Roberts (1985.)

Other demographic factors impacting on population structure

Brennan and Dyke found that both males and females who married and remained on Sanday were more likely than their single or emigrant counterparts to be first-born children and their average birth rank was higher. They also came from smaller families (Brennan and Dyke 1980). This is unsurprising: an eldest child from a small family with few competing brothers and sisters has a better chance of securing land or a dowry.

Sustained differential male and female emigration patterns will have an impact on population structure. This has not been specifically investigated in Orkney; however there is some evidence of excess female emigration in the nineteenth and early twentieth centuries, driven by lack of female employment opportunities in predominantly agricultural communities (Coull 1966). Male land inheritance may also contribute.

2.5 Conclusions

The history of Orkney over the last 250 years is marked by a long period of steep population decline, particularly in the remoter isles, as young people increasingly left Orkney in pursuit of better economic prospects elsewhere. Although there is evidence that this population decline has been stemmed since the 1970s, this has been the result of immigration, primarily from England, rather than through improved retention of the indigenous population. This all adds up to a picture of gradual isolate break-down.

The purpose of the present study is not to investigate levels of endogamy and inbreeding in Orkney in order to assess the evidence for isolate breakdown: to do so would require very careful sampling to ensure a representative cross-section of the population, which is beyond the scope of ORCADES. Orcadians were invited to volunteer for ORCADES if they had at least one grandparent born in the North Isles of Orkney. Out of the 1071 individuals satisfying this requirement who signed up for the study, those in the final 725 were chosen in order to maximise the range of inbreeding in the sample: both the half Orcadian group and those with detected inbreeding loops in their pedigrees are therefore likely to be over-represented, thus precluding extrapolation of these results to the wider Orcadian, or even North Isles, population. It is, nevertheless, interesting to view the evidence of inbreeding and endogamy in the ORCADES sample in the light of evidence of isolate breakdown. In this context, it is striking that evidence of inbreeding was detected in almost one quarter of the ORCADES sample and that over half of the sample had at least 3 grandparents born on the same isle or Orkney Mainland parish. Although the mean year of birth of the Half Orcadian group (1958) was later than that of either the mixed (1952) or endogamous (1951) groups, this difference was not significant. Furthermore, the estimates of inbreeding in the ORCADES sample are comparable to those found by both Brennan (1981) and Roberts (1979) in much earlier time periods. This study does not, then, provide strong evidence in support of isolate breakdown in Orkney. On the other hand, the mean birth year of the sample is 1952, so these findings relate primarily to patterns of endogamy in the middle third of the twentieth century. It seems unlikely that Orkney could have remained untouched by the accelerating pace of social change over the last half century.

Chapter 3: Runs of Homozygosity in European Populations

3.1 Introduction

Chapter 1 considers why measuring the effects of parental relatedness at an individual level is of epidemiological interest, highlights the deficiencies of using pedigree data to do this and describes different approaches using genomic data to quantify both autozygosity and homozygosity. Here a multipoint, observational approach to estimating autozygosity from genomic data is developed, which exploits the fact that autozygous genotypes are not evenly distributed throughout the genome, but are distributed in runs or tracts (Figure 1.1). This idea was first suggested by Broman and Weber, who proposed identifying autozygous segments from runs of consecutive homozygous markers (Broman and Weber 1999). There are three objectives to this study:

- To identify and describe ROH observable from high-density genome scan data in two isolated and two more cosmopolitan populations of European origin. The key study population is the ORCADES cohort described in chapter 2. Three additional populations are used for comparison: a representative Scottish comparison population (SOCCS) (Tenesa, Farrington et al. 2008), an isolate population from a Dalmatian island in Croatia (CROAS) (Campbell, Carothers et al. 2007) and the HapMap CEU founders from CEPH (northwest European-derived population from Utah, USA) (Frazer 2003).
- To investigate whether mean ROH statistics reflect differences in demographic history. Where possible, the study populations are sub-divided

according to levels of grandparental endogamy in order to see whether these differences are reflected in mean ROH statistics.

- To explore whether ROH can be used to provide an individual inbreeding coefficient which reliably reflects the genomic effects of recent parental relatedness. Using high quality pedigree information available for the ORCADES population, correlations are reported between F_{ped} and a genome-wide autozygosity measure derived from ROH (F_{ROH}) and these are compared with correlations between F_{ped} and 3 alternative genomic measures of homozygosity or autozygosity.

3.2 Subjects and Methods

3.2.1 The study populations

ORCADES received ethical approval from the appropriate research ethics committees in 2004 (Appendix 1). Data collection was carried out in Orkney between 2005 and 2007. 1019 Orcadian volunteers with at least one grandparent from the North Isles of Orkney gave informed consent and provided a blood sample. The mean age of the sample is 55 (dates of birth ranged from 1909 to 1988).

A Scottish comparison population was derived from the controls of the Scottish Colon Cancer Study (SOCCS) (Tenesa, Farrington et al. 2008). This consists of 984 subjects not known to have colon cancer matched by residential postal area and age to a series of incident cases of colorectal cancer. Subjects were resident throughout Scotland, with dates of birth ranging from 1921 to 1983 (mean 1952).

The CROAS sample consists of 849 Croatian individuals aged 18-93 sampled from the population of one island in 2003 - 2004 (Campbell, Carothers et al. 2007). Both SOCCS and CROAS were approved by the relevant ethics committees.

The CEU sample consists of 60 unrelated individuals from Utah, USA of northwest European ancestry, collected by the Centre d'Étude du Polymorphisme Humain (CEPH) in 1980 (2003).

3.2.2 Genotyping

Genotyping procedures for the SOCCS (Tenesa, Farrington et al. 2008), CROAS (Vitart, Rudan et al. 2008) and CEU (Frazer, Ballinger et al. 2007) samples are described elsewhere. All were genotyped using Illumina Infinium HumanHap300 platform (Illumina, San Diego). After extraction of genomic DNA from whole blood using Nucleon kits (Tepnel, Manchester), 758 ORCADES samples were genotyped according to the manufacturer's instructions on the Illumina Infinium HumanHap300v2 platform (Illumina, San Diego). Analysis of the raw data was done using BeadStudio software with the recommended parameters for the Infinium assay using the genotype cluster files provided by Illumina.

Individuals with less than 95% call rate were removed, as were SNPs with more than 10% missing. SNPs failing HWE at a threshold of $p = 0.0001$ were removed, as this may reflect poor genotyping of these SNPs. IBD sharing between all first and second degree relative pairs was assessed using the *Genome* program in PLINK, available at http://pngu.mgh.harvard.edu/purcell/plink/ (Purcell 2007; Purcell, Neale

et al. 2007) and individuals falling outside expected ranges were removed from the study. Sex checking was performed using PLINK and individuals with discordant pedigree and genomic data were removed. On completion of data cleaning and quality control procedures, 725 individuals and 316, 364 autosomal SNPs remained. 45% are male.

A consensus SNP panel was then created, using only those markers that satisfied these QC criteria in all 4 populations, leaving a final sample of 289,738 autosomal SNPs and 2618 individuals (60 CEU, 725 ORCADES, 849 CROAS and 984 SOCCS).

3.2.3 F_{ped} estimates

The pedigrees of all individuals in the ORCADES sample were traced back for as many generations as possible in all ancestral lineages, using official birth, marriage, death and census records held by the General Register Office for Scotland in Edinburgh. This involved tracing the records and recording the details of over 12,000 individuals. F_{ped} was calculated for each individual using Wright's path method (Wright 1922). Full details are given in chapter 2.

Limited pedigree information is available for the CROAS data set. Very few individuals had complete pedigrees to three ancestral generations, the minimum required to identify the offspring of first cousins, so it was not possible to derive individual estimates of F_{ped}. Grandparental information was, however, fairly complete, so it was possible to categorise individuals in terms of grandparental

birthplace and therefore to assess the association between endogamy and ROH in this sample.

No pedigree information is available for the SOCCS data set; however data were analysed according to the rurality of subjects' residential address to investigate whether there appears to be any association between remote rurality and autozygosity in Scotland. Two measures of rurality were used: firstly, subjects were classified on the basis of residential postcodes according to the Scottish Household Survey 6-fold urban/rural classification (table 3.1) (SE 2004), available at http://www.scotland.gov.uk/Publications/2004/. Secondly, the analysis was repeated, collapsing categories 1 – 3 into a broad urban group and categories 4 – 6 into a broad rural group. The small number of island residents was separated out into a third group. Data on grandparental country of birth were available for a subset of this sample. Mean homozygosity and autozygosity measured in various ways (see below) of those with 4 Scottish-born grandparents were compared with those who had at least one grandparent born outside Scotland.

Table 3.1: Scottish Executive Urban Rural Classification, 2003 - 2004

Category	Description
1	Large Urban Areas - Settlements of over 125,000 people.
2	Other Urban Areas - Settlements of 10,000 to 125,000 people.
3	Accessible Small Towns - Settlements of between 3,000 and 10,000 people and within 30 minutes drive of a settlement of 10,000 or more.
4	Remote Small Towns - Settlements of between 3,000 and 10,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more.
5	Accessible Rural - Settlements of less than 3,000 people and within 30 minutes drive of a settlement of 10,000 or more.
6	Remote Rural - Settlements of less than 3,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more.

3.2.4 Runs of Homozygosity

ROH were identified using the Runs of Homozygosity programme implemented in PLINK version 1.0 (Purcell; Purcell, Neale et al. 2007). This slides a moving window of 5000 kb (minimum 50 SNPs) across the genome to detect long contiguous runs of homozygous genotypes. An occasional genotyping error or missing genotype occurring in an otherwise unbroken homozygous segment could result in the under-estimation of ROH. To address this, the program allows one heterozygous and 5 missing calls per window.

The aim of this analysis is to identify and quantify ROH of different lengths in order to assess the extent to which they result from parental relatedness and population isolation and the extent to which they do not. Ideally, the aim would be to identify all ROH, regardless of how short; however in reality, limitations on the length of ROH it is possible to identify are set by the density of SNP panel used in the

analysis. This is explored in more detail in chapter 4, where it is shown that the 300K panel under-estimates the proportion of the genome in ROH shorter than 1.5 Mb compared with estimates made using a denser SNP panel. For the present analysis, the minimum length of ROH is set at 500 kb, although most analyses are conducted using the more reliable 1.5 Mb threshold. The reason for retaining the 500 kb limit is that it gives an (albeit under-estimated) indication of the prevalence of ROH of intermediate length which result from the inheritance through both parents of haplotypes that are at high frequency in the population. The ideal would be to measure ROH resulting from all LD, which typically extends up to about 100 kb in the human genome (Abecasis, Noguchi et al. 2001; Reich, Cargill et al. 2001; Wall and Pritchard 2003; Abecasis, Ghosh et al. 2005); however this is not possible with a 300K SNP panel. All empirical studies have identified a few much longer stretches of LD in the human genome, measuring up to several hundred kb in length (Wall and Pritchard 2003), which may result in the occurrence of longer ROH in outbred individuals. Such longer stretches can be identified with the 300K panel, with the caveat that ROH shorter than 1.5 Mb are likely to be under-estimated.

PLINK also requires the specification of a minimum number of consecutive homozygous SNPs constituting a ROH. The greater the number of consecutive homozygous genotype calls, the more likely the ROH represents a true homozygous segment. With, for example, only 3 consecutive homozygous genotypes, there would be a very high probability that these 3 could be homozygous by chance alone (on the basis of allele frequencies) and that the intervening, unobserved chromosomal stretches could be heterozygous. The level of confidence that a ROH

measuring 500 kb and containing 30 SNPs represented a true homozygous segment would obviously be higher than the level of confidence in a ROH of the same length but containing only 5 SNPs. The density of SNP coverage in the study panel is 9.23 kb/SNP, which means that, on average, a 500 kb stretch will contain 54 SNPs. There is, however, a great deal of variation in the density of SNP coverage across the genome, so for this reason, the minimum number of SNPs constituting a ROH was set at 25. This was felt to be high enough to minimize the erroneous identification of non-homozygous segments as ROH, whilst also being able to identify ROH in sparsely covered genomic regions. Two further parameters were included: tracts with a mean tract density > 50 kb/SNP were excluded, and the maximum gap between two consecutive homozygous SNPs was set at 100 kb.

In order to exclude the possibility that apparent ROH are in fact regions of hemizygous deletion, an analysis of deletions was carried out in the Orkney data set. An Objective Bayes' Hidden Markov model, as employed in QuantiSNP v. 1.0, was used to identify heterozygous deletions with a sliding window of 2 Mb over the genome and 25 iterations. This work was undertaken by Rehab Abdel-Rahman. All the samples were corrected for genome GC-content prior to copy number inference to ensure the variation of the observed $\log_2$ R ratio is not attributed to the specific regional GC-content (Marioni, Thorne et al. 2007). All heterozygous deletions with estimated Bayes' factor ≥ 10 were included in the downstream analysis to ensure a low false negative rate as reported in Colella et al, 2007 (Colella, Yau et al. 2007). A custom Perl script was developed to compare the identified heterozygous deletions and ROH. All deletions overlapping with ROH were identified. Where deletions

covered the entire length of the ROH or where less than 0.5 MB of the tract remained after taking account of the deletion, the ROH was removed from the analysis (*personal communication*, Rehab Abdel-Rahman). Because the CROAS, CEU and SOCCS data sets were uncorrected for deletions, uncorrected ORCADES data are shown where there are population comparisons. Analyses using only the ORCADES data set use data corrected for deletions.

3.2.5 F_{ROH}

A genomic measure of individual autozygosity (F_{ROH}) was derived, defined as the proportion of the autosomal genome in runs of homozygosity above a specified length threshold:

$$F_{ROH} = \sum L_{ROH} / L$$

where $\sum L_{ROH}$ is the total length of all an individual's ROH above a specified minimum length and L is the length of the autosomal genome covered by SNPs, excluding the centromeres. The centromeres are excluded because they are long genomic stretches devoid of SNPs and including them might inflate estimates of autozygosity if both flanking SNPs are homozygous. The length of the autosomal genome covered by the consensus panel of SNPs is 2,673,768 kb. Individual and population mean values of F_{ROH} are shown for a range of different ROH length thresholds.

3.2.6 Alternative genomic measures of autozygosity or homozygosity

F_{ROH} statistics are compared with 3 other genomic measures. These are described more fully in chapter one, and are summarized briefly here. Multi-locus

heterozygosity (MLH) (Charpentier, Setchell et al. 2005) is simply a measure of the proportion of typed genotypes that are heterozygous. To express this as a measure of homozygosity rather than heterozygosity, the statistic used here is 1-MLH, termed H_{pn} throughout this thesis. F_{plink} is a genomic estimate of autozygosity implemented in PLINK (Purcell, Neale et al. 2007). This uses expected genome heterozygosity to control for background homozygosity. F_{plink} is closely related to a measure of excess homozygous genotypes (i.e. observed minus expected homozygous genotypes, termed here H_{ex}, with expected homozygous genotypes estimated from sample allele frequencies on the basis of Hardy Weinberg expectation). H_{ex} is simply the numerator of $F_{plink.}$

3.2.7 Statistical analysis

For statistical analyses, the ORCADES sample was split into endogamous Orcadians, defined as those with at least 3 grandparents born on the same isle within Orkney, typically ~ 10 km^2 in size with a population of 50 – 500 (n = 390); mixed Orcadians, defined as those with at least 3 grandparents born in Orkney but not on the same island - i.e. from an area over 500 km^2 with a population of ~ 20,000 (n = 286); and half Orcadians, defined as those with 1 pair of Orcadian and 1 pair of Scottish-mainland-born grandparents (n = 49). Although pedigree information is not available to assess whether the parents of half Orcadian subjects are related beyond 5 generations in the past, it is reasonable to assume that they are likely to be unrelated for at least 10-12 generations. It is known that there was major Scottish immigration to Orkney in the 15^{th} and 16^{th} centuries, before 10 – 12 generations ago. Although Scottish immigration has certainly occurred sporadically since then, rates have been

low. An analysis of the area of origin of the Scottish parents of our half Orcadian subjects shows that they came from all over Scotland: we found no evidence for strong Orcadian connections with any specific Scottish settlement, which might have increased the chances of parental relatedness in this group. Furthermore, the surnames of the ancestors of the Orcadian parents of this group were markedly different from those of the ancestors of the non-Orcadian Scottish parents.

The CROAS sample was split into endogamous Dalmatian, defined as those with all 4 grandparents born in the same village – i.e. from a 1 km^2 area, with a population of < 2000 (n = 431); mixed Dalmatian, defined as those with all 4 grandparents born on the same island but not in the same village – i.e. from a 90 km^2 area with a population of 3,600 (n = 221); and Croatian, defined as residents of the island with grandparents born elsewhere in Croatia (n = 197). The CEU and SOCCS samples were not sub-divided, with one exception: a subset of the SOCCS sample, consisting of individuals for whom grandparental country of birth was available, was used in some analyses.

All calculations were performed using SPSS and Excel software. The proportions of each sub-population with ROH measuring less than 1, 1.5, and 2 Mb were calculated. All subjects in all sub-populations had ROH shorter than 1.5 Mb. Sub-populations start to become differentiated from each other for ROH > 1.5 Mb, with the effects of endogamy on ROH starting to emerge above this threshold. Unless otherwise specified, all analyses exploring the effects of endogamy and parental relatedness on ROH therefore define a ROH as measuring ≥ 1.5 Mb.

Sub-population means were calculated for the total length of ROH per individual. Number of ROH was plotted against total length of ROH per individual for each sub-population.

The correlation between F_{ped} and F_{ROH} was calculated using a subset of 249 individuals from the ORCADES sample. This subset had at least 2 grandparents on the same side of the family born in Orkney and no grandparents born outside Scotland. They were either the offspring of consanguineous parents (parents related as 2nd cousins or closer) or they were the offspring of non-consanguineous people for whom it was possible to establish pedigrees for at least 5 ancestral generations in all Orcadian ancestral lineages, or 4 ancestral generations in non-Orcadian ancestral lineages. The reason for the difference in parameters between Orcadian and non-Orcadian lineages is that it is much more difficult to trace ancestry in mainland Scotland than ancestry in Orkney. Any inbreeding loops detected beyond the 4 or 5 ancestral generation limit were disregarded, as pedigrees were very incomplete beyond this limit.

Correlations were also calculated between F_{ROH}, F_{ped} and the three other measures of autozygosity or homozygosity described above (H_{pn}, F_{plink} and H_{ex}).

3.2.8 Prevalence and genomic location of ROH in different sub-populations

Next, the hypothesis that ROH in outbred individuals tend to cluster in the same genomic locations, whereas those present in the offspring of related parents tend to

be more randomly distributed across the autosomal genome was explored. The location of ROH in 3 groups was compared: the half Orcadian group, consisting of all half Orcadians with at least one ROH measuring ≥ 1.5 Mb (n = 46); an offspring of cousins group, which was constructed by taking all individuals in the ORCADES sample with parents related as 3rd cousins or closer and then choosing the 20 with the greatest total length of ROH; and a control population derived from the SOCCS sample. Because some individuals in the SOCCS sample have long ROH which may be indicative of parental relatedness, the control sample was restricted to those with no more than 8 ROH, totalling no more than 17 Mb. These are the maximum values in the half Orcadian group, the members of which are known to be the offspring of unrelated parents. There were 943 individuals in the control group. ROH measuring at least 1.5 Mb in all three groups were compared. ROH in the control group overlapping by at least 0.5 Mb with ROH in either ORCADES group were counted. The number of control overlaps per ROH (and per Mb of ROH) in the half Orcadian group was compared with those in the offspring of cousins group.

Next, the question of whether ROH in half Orcadians occurred in regions of lower than average recombination was investigated. Based on sex-averaged mean recombination rates per Mb derived from the deCODE genetic map, the UCSC Genome Browser was used, available at http://genome.ucsc.edu/cgi-bin/hgGateway (Kent, Sugnet et al. 2002) to calculate the mean recombination rate of all complete Mb of ROH in the half Orcadian sample.

3.3 Results

3.3.1 Copy number variation

224 deletions were detected which overlapped with ROH (median length of deletion 995 kb). Overlapping deletions were detected in 57 individuals (7.6% of sample). (*personal communication*, Rehab Abdel-Rahman). After removal of these overlaps from the sample, and removal of the entire affected ROH if less than 0.5 Mb remained, ROH statistics were recalculated. There was no significant difference between results before and after correction for deletion for mean total length of ROH (correcting for deletions reduced this by less than 0.3% in the sample as a whole) or mean number of ROH (reduced by 0.02%). Furthermore, no significant differences were found when data were analysed by sub-population and when different length parameters were used to define ROH. This provides strong evidence that the ROH identified are true homozygous tracts and not hemizgyous deletions.

3.3.2 Urban/rural analysis of SOCCS sample

No difference was found in mean F_{ROH}, H_{pn}, F_{plink} or H_{ex} between those living in rural and urban areas of Scotland, regardless of whether the analysis used a detailed 6-category classification from large urban to remote rural or a broader, 3 category classification (tables 3.2 and 3.3). For ease of interpretation, F_{ROH} statistics are expressed here as percentages rather than proportions (i.e. F_{ROH} is here defined as the percentage of the typed autosomal genome in ROH).

Table 3.2: Mean (95% confidence interval) F_{ROH} (expressed as a percentage), H_{pn}, F_{plink} and H_{ex} of SOCCS sample by urban rural classification (6 categories)

Category	Measure	N	Mean	SE	95% Confidence Interval
Large urban	**$F_{ROH0.5}$ (%)**	351	3.09	0.019	3.05 to 3.12
Other urban		314	3.12	0.018	3.09 to 3.16
Accessible small town		75	3.15	0.087	2.98 to 3.32
Remote small town		49	3.07	0.045	2.98 to 3.16
Accessible rural		129	3.10	0.027	3.05 to 3.15
Remote rural		66	3.13	0.043	3.05 to 3.22
Large urban	**$F_{ROH1.5}$ (%)**	351	0.259	0.0097	0.240 to 0.278
Other urban		314	0.272	0.0120	0.248 to 0.295
Accessible small town		75	0.349	0.0800	0.192 to 0.506
Remote small town		49	0.253	0.0209	0.212 to 0.294
Accessible rural		129	0.257	0.0146	0.228 to 0.286
Remote rural		66	0.291	0.0254	0.241 to 0.340
Large urban	**F_{ROH5} (%)**	351	0.0127	0.004114	0.0047 to 0.0208
Other urban		314	0.0165	0.005984	0.0047 to 0.0282
Accessible small town		75	0.0864	0.060589	0 to 0.2051
Remote small town		49	0.0049	0.004862	0 to 0.0144
Accessible rural		129	0.0105	0.004862	0.0009 to 0.0200
Remote rural		66	0.0153	0.010846	0 to 0.0366
Large urban	**H_{pn}**	351	0.6507	0.0001	0.6505, 0.6510
Other urban		314	0.6508	0.0001	0.6506, 0.6511
Accessible small town		75	0.6509	0.0003	0.6503, 0.6516
Remote small town		49	0.6507	0.0003	0.6501, 0.6513
Accessible rural		129	0.6508	0.0002	0.6505, 0.6511
Remote rural		66	0.6510	0.0002	0.6506, 0.6514
Large urban	**F_{plink}**	351	-0.00010	0.00035	-0.00078, 0.00059
Other urban		314	0.00018	0.00034	-0.00049, 0.00084
Accessible small town		75	0.00053	0.00094	-0.00131, 0.00236
Remote small town		49	-0.00014	0.00082	-0.00174, 0.00146
Accessible rural		129	0.00014	0.00047	-0.00078, 0.00105
Remote rural		66	0.00072	0.00063	-0.00052, 0.00195
Large urban	**H_{ex}**	351	-13	35	-82, 56
Other urban		314	15	34	-53, 82
Accessible small town		75	50	95	-135, 236
Remote small town		49	-21	83	-184, 143
Accessible rural		129	12	47	-81, 105
Remote rural		66	70	64	-55, 195

Table 3.3: Mean (95% confidence interval) F_{ROH} (expressed as a percentage), H_{pn}, F_{plink} and H_{ex} of SOCCS sample by urban rural classification (3 categories)

Category	Measure	N	Mean	SE	95% confidence interval
Urban	$F_{ROH0.5}$ (%)	742	3.11	0.015	3.08 to 3.14
Rural		226	3.10	0.021	3.06 to 3.14
Island		16	3.19	0.087	3.01 to 3.36
Urban	$F_{ROH1.5}$ (%)	742	0.273	0.0105	0.252 to 0.294
Rural		226	0.263	0.0116	0.240 to 0.286
Island		16	0.304	0.0516	0.203 to 0.405
Urban	F_{ROH5} (%)	742	0.0217	0.00711	0.0078 to 0.0356
Rural		226	0.0082	0.00337	0.0016 to 0.0148
Island		16	0.0475	0.03478	0 to 0.1157
Urban	H_{pn}	742	0.65079	0.000083	0.65062, 0.65095
Rural		226	0.65085	0.000125	0.6506, 0.65109
Island		16	0.65087	0.000406	0.65008, 0.65167
Urban	F_{plink}	742	0.000075	0.000238	-0.000392, 0.000542
Rural		226	0.000259	0.000358	-0.000443, 0.000960
Island		16	0.000331	0.001163	-0.001949, 0.002611
Urban	H_{ex}	742	4	24	-43, 52
Rural		226	23	36	-48, 94
Island		16	32	118	-200, 263

3.3.3 The effect of stochastic variation on individual autozygosity

On average, the difference between full sibling pairs in the total length of ROH was 10.3 Mb. However the distribution is skewed with half of all individuals having less than 5 Mb difference, yet some 7% differing by more than 30 Mb. The greatest difference between sib pairs was 91 Mb, or 3.4% of the autosomal genome (paternity was confirmed from patterns of genomic sharing in all cases).

3.3.4 Effects of population isolation and endogamy on length and number of ROH

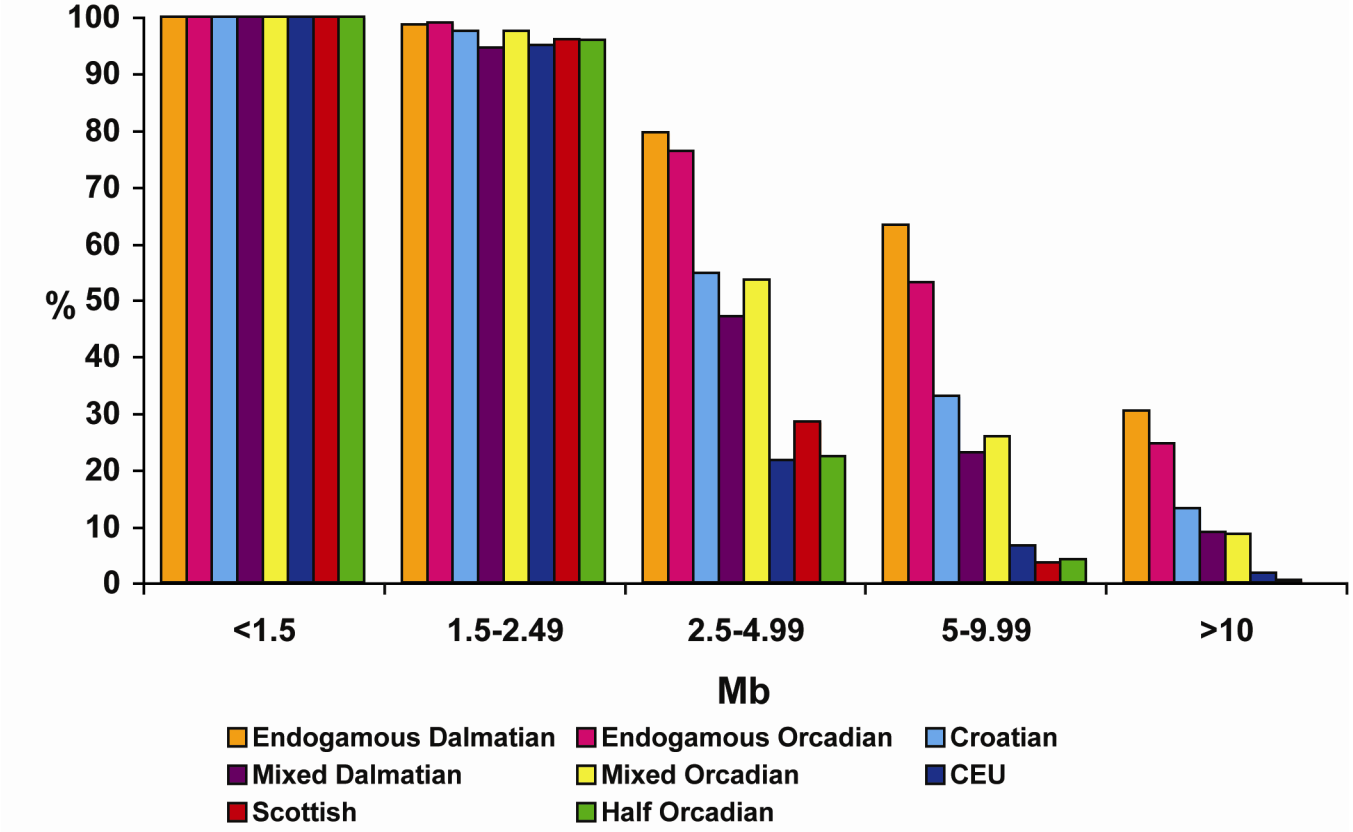
The proportions of sub-populations with ROH of a given length are shown in figure 3.1. All individuals in all populations have ROH measuring less than 1.5 Mb.

Taking the populations as a whole, on average a significantly greater proportion of the autosomal genome of the ORCADES sample are in ROH measuring 0.5 – 1.5 Mb (77.7 Mb; 95 % confidence interval 77.1 to 78.2) than is the case for either the CROAS (73.2 Mb; 95% confidence interval 72.7 to 73.7) the SOCCS (75.8 Mb; 95% confidence interval 75.3 to 76.3) or the CEU (74.1 Mb; 95% confidence interval 72.3 to 75.8) samples. There are no significant differences between groups within populations, however, which suggests that this reflects population differences in genetic diversity or LD of ancient origin rather than the effects of more recent endogamy or population isolation (although it should be remembered that estimates of these shorter ROH are less reliable than estimates of ROH longer than 1.5 Mb).

For ROH above 1.5 Mb, 3 distinct groupings emerge which are clearly related to endogamy and isolation: a greater proportion of the endogamous Dalmatian and Orcadian samples than of the other samples have long ROH (28% have ROH > 10 Mb); only a small proportion of the CEU, SOCCS and half Orcadian samples have long ROH (0.5% >10 Mb), with the Croatian and mixed Dalmatian and Orcadian samples in between (10% > 10 Mb).

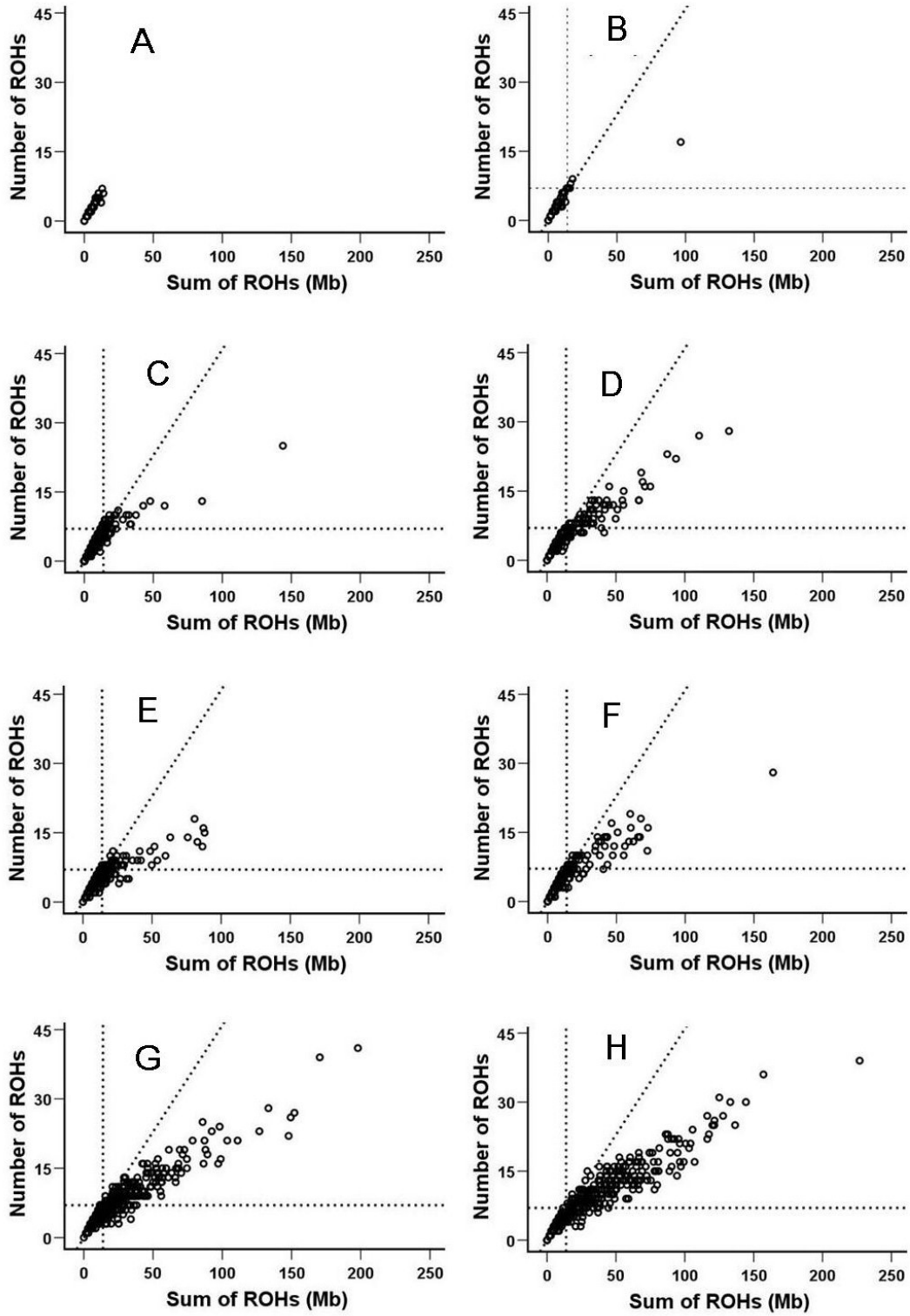
Forty-nine individuals had no ROH longer than 1.5 Mb. This included at least one individual from each sub-population, although they were predominantly from the half Orcadian, SOCCS and CEU samples. The shortest sum of ROH across all the samples was found in an individual in the SOCCS sample, who had ROH longer than 0.5 Mb covering only 1.5% of the autosomes (39 Mb). This compares with a mean of 3.5% across all the populations (93 Mb).

Figure 3.1: Proportion of sub-populations with one or more ROH of a given length



The number of ROH longer than 1.5 Mb per individual, plotted against the total length of those ROH, is shown in figure 3.2 for each group. The half Orcadian group is used as a reference, as we know that these individuals are the offspring of unrelated parents. Reference lines are shown on all graphs for the maximum number of ROH, the maximum total length of ROH and the line of best fit for the half Orcadian group. Compared with the half Orcadian group, all other groups have a greater variance in the number and sum of ROH and contain individuals with more and longer ROH. Again, the same three groupings are apparent. Data points for the half Orcadian, SOCCS and CEU samples are generally narrowly distributed along both axes, indicating that these individuals have few, relatively short ROH. The two endogamous samples are much more widely spread along both axes, reflecting the presence of many, much longer ROH. The Croatian, mixed Orcadian and mixed Dalmatian groups are intermediate, reflecting the fact that these less carefully specified groups are probably made up of individuals with a mixture of ancestries, from the outbred to the very endogamous. The percentage of each group with more and longer ROH than the maximum for the half Orcadians was calculated. Again, the SOCCS (5%) and CEU (8%) groups differed least and the endogamous Dalmatians (64%) and Orcadians (54%) differed most from the half Orcadians. The Croatians (33%) and mixed Dalmatians (26%) and Orcadians (23%) were intermediate.

Figure 3.2: Number of ROH compared to total length of ROH



Panels: A - Half Orcadian, B - CEU, C - SOCCS, D - Croatian, E - Mixed Orcadian, F - Mixed Dalmatian, G - Endogamous Orcadian, H - Endogamous Dalmatian.

Population sub-group means and 95% confidence intervals for the ORCADES and CROAS samples were calculated for F_{ROH}, F_{plink}, H_{pn} and H_{ex}, in order to investigate whether there were significant differences between the sub-groups of each population. Data are shown in table 3.4. In the ORCADES sample, mean F_{ROH} in the endogamous group is significantly higher than mean F_{ROH} in the mixed group, which is in turn significantly higher than mean F_{ROH} in the half Orcadian group, regardless of the ROH length cut-off used. This is illustrated graphically in figures 3.3 and 3.4, which show that the 95% confidence intervals of each sub-group do not overlap. There is, however, no significant difference between the mixed and half Orcadian groups for F_{plink}, H_{pn} or H_{ex}, although the endogamous group remains significantly higher than the other two groups according to these three measures. In the CROAS sample there is no overlap between the 95% confidence interval for the endogamous group and the other two groups on any of the six measures; however the mixed and Croatian groups have overlapping confidence intervals regardless of the measure used, indicating that there is no significant difference between these two categories. The Croatian sub-group is a group of settlers for whom grandparental data were not available, so they are not as precisely specified in terms of endogamous ancestry as the other two CROAS sub-groups. One further comparison between sub-groups using these six measures was performed: grandparental country of birth was available for a subset (n = 426) of the SOCCS sample. On average, those with 4 Scottish-born grandparents (n = 254) had slightly greater F_{ROH} than those with at least one grandparent born outside Scotland, but differences were not significant, regardless of the ROH length cut-off used. Differences between the two groups were, however, significant when analysed using F_{plink}, H_{pn} and H_{ex} (table 3.5).

Table 3.4: Mean (95% confidence interval) F_{ROH}, H_{pn}, F_{plink} and H_{ex} of ORCADES and CROAS sub-populations

Population	Sub-population	Measure	N	Mean	SE	95% Confidence Interval
CROAS	**Croatian settler**	H_{pn}	197	0.655	0.000313	0.654, 0.656
	Mixed		221	0.654	0.000292	0.654, 0.655
	Endogamous		431	0.657	0.000255	0.657, 0.658
ORCADES	**Endogamous**		390	0.655	0.000251	0.655, 0.656
	Half		49	0.652	0.00045	0.651, 0.653
	Mixed		286	0.653	0.000198	0.652, 0.653
CROAS	**Croatian settler**	F_{plink}	197	0.0066	0.00089	0.0048, 0.0083
	Mixed		221	0.0045	0.000829	0.0028, 0.0061
	Endogamous		431	0.0135	0.000728	0.0121, 0.0149
ORCADES	**Endogamous**		390	0.0009	0.000721	-0.0005, 0.0024
	Half		49	-0.0077	0.001273	-0.0102, -0.0052
	Mixed		286	-0.0056	0.000565	-0.0067, -0.0045
CROAS	**Croatian settler**	H_{ex}	197	649	88	476, 822
	Mixed		221	441	82	279, 602
	Endogamous		431	1344	73	1202, 1487
ORCADES	**Endogamous**		390	93	71	-47, 233
	Half		49	-768	126	-1014, -521
	Mixed		286	-553	56	-663, -444
CROAS	**Croatian settler**	$F_{ROH0.5}$	197	3.44	0.060	3.32 to 3.55
	Mixed		221	3.27	0.050	3.17 to 3.37
	Endogamous		431	4.06	0.056	3.95 to 4.17
ORCADES	**Endogamous**		390	3.95	0.054	3.85 to 4.06
	Half		49	3.14	0.045	3.06 to 3.23
	Mixed		286	3.44	0.034	3.37 to 3.50
CROAS	**Croatian settler**	$F_{ROH1.5}$	197	0.69	0.054	0.58 to 0.80
	Mixed		221	0.56	0.047	0.47 to 0.65
	Endogamous		431	1.32	0.054	1.21 to 1.42
ORCADES	**Endogamous**		390	1.06	0.050	0.96 to 1.15
	Half		49	0.22	0.019	0.18 to 0.26
	Mixed		286	0.53	0.031	0.47 to 0.59
CROAS	**Croatian settler**	F_{ROH5}	197	0.25	0.034	0.18 to 0.32
	Mixed		221	0.18	0.032	0.11 to 0.24
	Endogamous		431	0.64	0.039	0.56 to 0.72
ORCADES	**Endogamous**		390	0.46	0.036	0.39 to 0.53
	Half		49	0.0097	0.007	0 to 0.02
	Mixed		286	0.16	0.024	0.12 to 0.21

Table 3.5: Differences in estimates of F_{ROH}, H_{pn}, F_{plink} and H_{ex} between those with 4 Scottish-born grandparents and those with fewer than 4 Scottish-born grandparents (SOCCS sample)

Measure	Category	Mean	N	SE	P
$F_{ROH0.5}$	4 Scottish-born grandparents	0.0312	254	0.00024	0.27
	< 4 Scottish-born grandparents	0.0308	172	0.00030	
$F_{ROH1.5}$	4 Scottish-born grandparents	0.0028	254	0.00016	0.63
	< 4 Scottish-born grandparents	0.0027	172	0.00021	
F_{ROH5}	4 Scottish-born grandparents	0.000264	254	0.00008	0.97
	< 4 Scottish-born grandparents	0.000258	172	0.00015	
H_{pn}	4 Scottish-born grandparents	0.65107	254	0.00013	0.01
	< 4 Scottish-born grandparents	0.65055	172	0.00016	
F_{plink}	4 Scottish-born grandparents	0.0009	254	0.00037	0.01
	< 4 Scottish-born grandparents	-0.00062	172	0.00045	
H_{ex}	4 Scottish-born grandparents	87.14	254	37.83	0.01
	< 4 Scottish-born grandparents	-62.63	172	45.63	

Figure 3.3: Mean (95% confidence interval) F_{ROH}, H_{pn}, F_{plink} and H_{ex} for ORCADES sub-populations

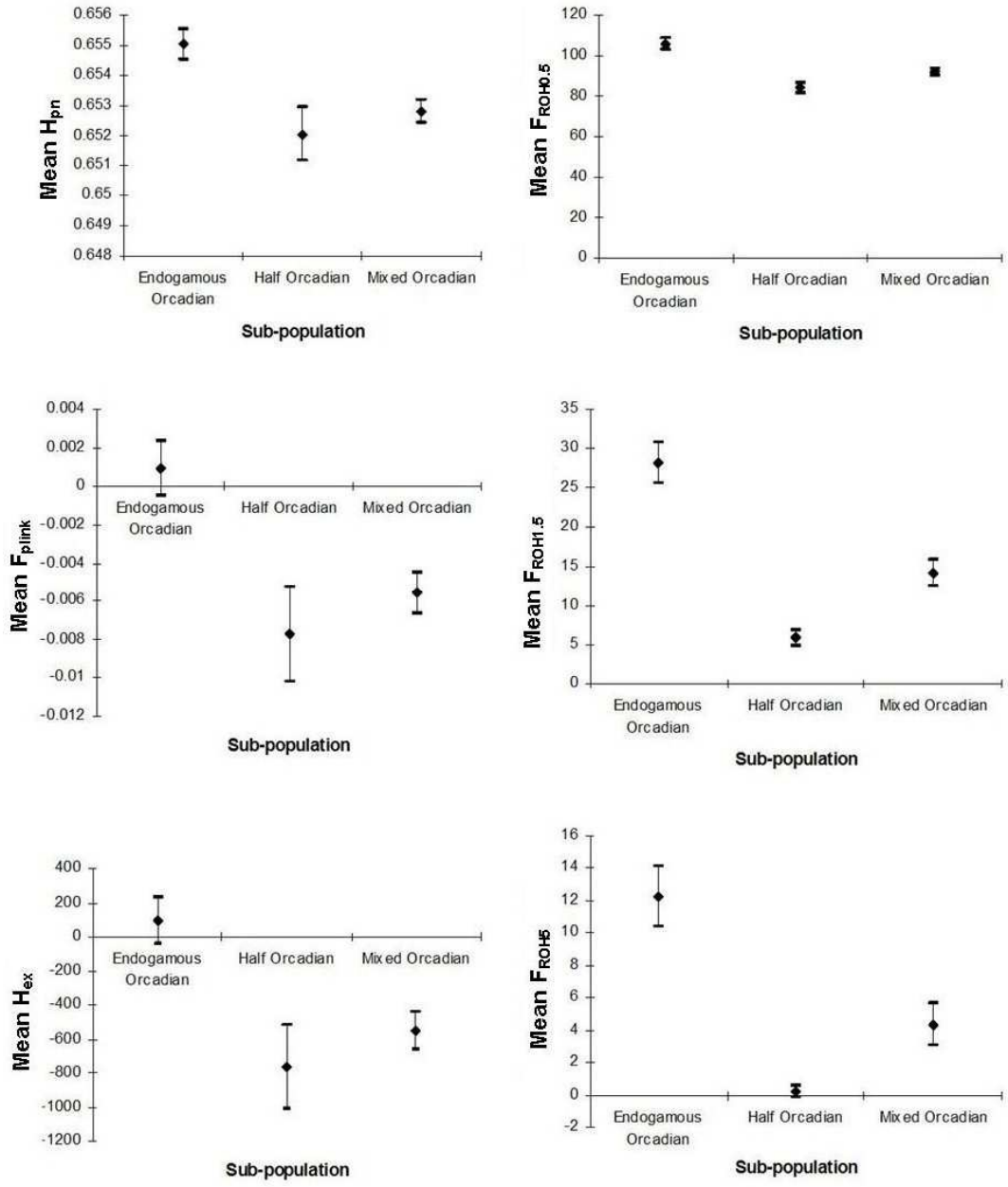
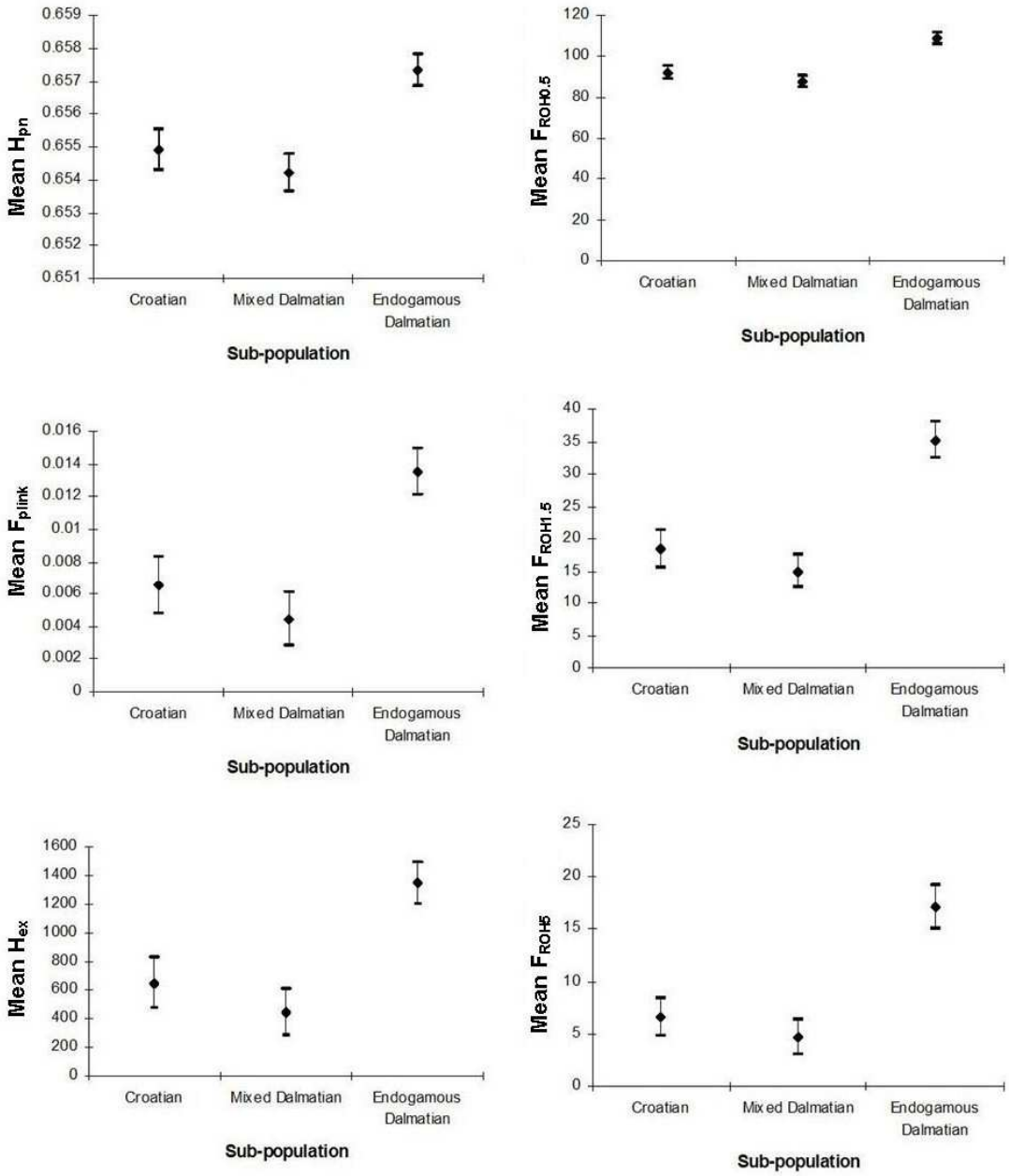
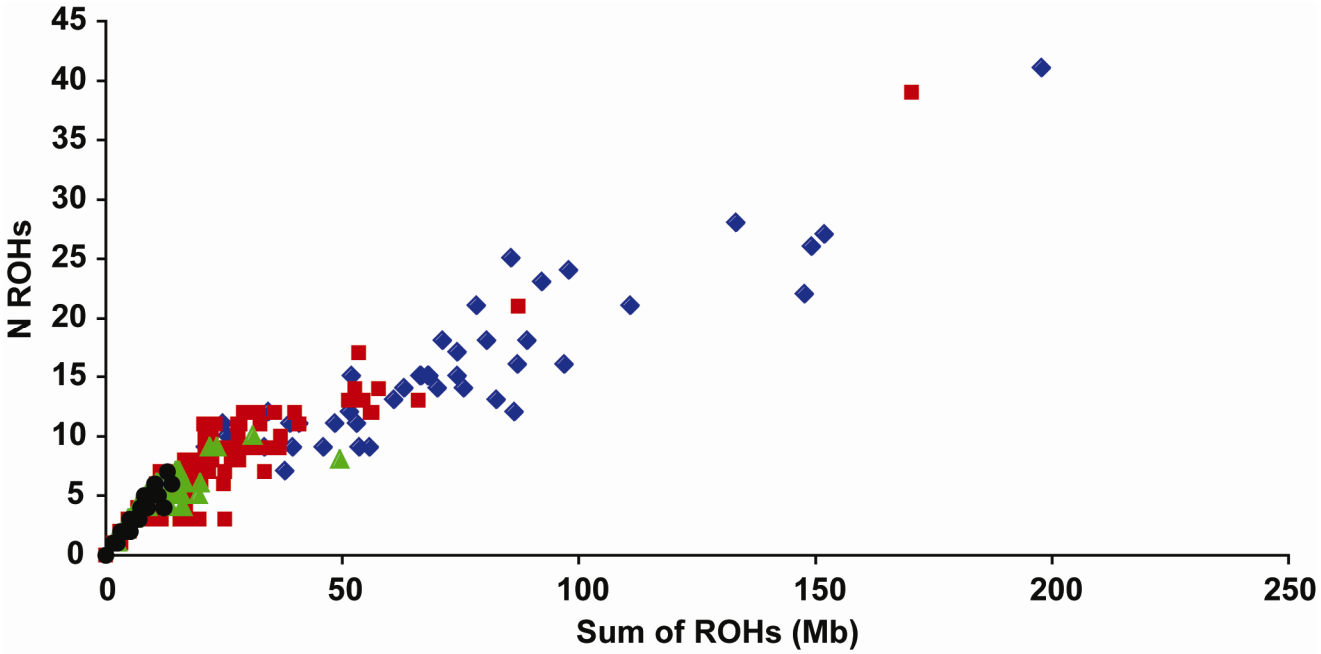


Figure 3.4: Mean (95% confidence interval) F_{ROH}, H_{pn}, F_{plink} and H_{ex} for CROAS sub-populations



The effect of different degrees of parental relatedness on the sum and number of ROH is shown in figure 3.5 for the 249 individuals in the Orkney sample with good pedigree information. Although a trend for increasing number and total length of ROH is evident from the half-Orcadian through the mixed to the endogamous and cousin offspring sub-groups, there is considerable overlap between groups.

Figure 3.5: Effect of endogamy on sum and number of ROH



Offspring of 1st or 2nd cousins are shown in blue; Endogamous Orcadians who are not the offspring of 1st or 2nd cousins are shown in red; Mixed Orcadians are shown in green and Half Orcadians are shown in black.

3.3.5 Comparison of F_{ped} and F_{ROH}

A subset of 249 individuals from the ORCADES sample with complete and reliable pedigree data were used to compare F_{ped} and F_{ROH}. The mean (standard error) F_{ped} of the sample is 0.0038 (0.0005), approximately equivalent to a parental relationship of third cousins. Mean F_{ped} values for ORCADES sub-populations are shown in Table 3.6. These vary from 0.02 for the offspring of 1st or 2nd cousins, to 0.0002 (equivalent to a parental relationship of 5th cousins) in the mixed Orcadian group. Mean F_{ped} values are compared with mean F_{ROH} values for a range of minimum length thresholds. The mean value of $F_{ROH\ 5}$ (ie using a minimum length threshold of 5 Mb) is closest to that of F_{ped}, whilst $F_{ROH\ 0.5}$ (ie using a minimum length threshold of 0.5 Mb) is an order of magnitude higher. This suggests that a shared maternal and paternal ancestor in the preceding 5 generations results predominantly in ROH longer than 5 Mb. It is clear from the half Orcadian group, whose parents do not share a common ancestor for at least 5 and probably at least 10 ancestral generations, that ROH measuring less than 3 or 4 Mb are not uncommon in the absence of parental relatedness. On average, these individuals have over 3% (84 Mb) of their autosomal genome in ROH over 0.5 Mb long and 0.2% (almost 6 Mb) in ROH longer than 1.5 Mb.

Table 3.6: Mean values of F_{ped} and F_{ROH} for ORCADES sub-populations (n = 249)

Orkney sub-population	N	Mean (SE) F_{ped}	Equivalent parental cousin relationship (single loop)	Mean (SE) $F_{ROH0.5}$	Mean (SE) $F_{ROH1.5}$	Mean (SE) F_{ROH5}
Offspring of 1st or 2nd cousins	42	0.0182 (0.0014)	2nd cousin	0.0569 (0.0024)	0.0271 (0.0022)	0.0169 (0.0017)
Endogamous Orcadian	114	0.0015 (0.0004)	3rd – 4th cousin	0.0379 (0.0008)	0.0087 (0.0007)	0.003 (0.0004)
Mixed Orcadian	44	0.0002 (0.0001)	5th cousin	0.033 (0.0006)	0.0046 (0.0005)	0.0012 (0.0004)
Half Orcadian	49	0	None	0.0315 (0.0004)	0.0021 (0.0002)	0.0001 (0.00007)
Total	249	0.0038 (0.0005)	3rd cousin	0.039 (0.0008)	0.0098 (0.0007)	0.0045 (0.0005)

3.3.6 Correlation between F_{ROH}, F_{ped}, F_{plink}, H_{pn} and H_{ex}

The total sample was used to examine correlations between different genetic estimates of autozygosity and homozygosity. Allele frequencies for F_{plink} and H_{ex} were estimated by naïve counting in all individuals, as implemented in PLINK ((Purcell 2007). Results are shown in table 3.7. F_{plink} and H_{ex} are almost perfectly correlated and F_{plink} and H_{pn} are highly correlated ($r = 0.938$). All 3 F_{ROH} measures correlate significantly more strongly with H_{pn} than with either H_{ex} or F_{plink}. Increasing the minimum ROH length threshold weakens correlations between F_{ROH} and the other 3 measures. F_{ROH5} is significantly more weakly correlated with all three alternative measures than are either $F_{ROH1.5}$ or $F_{ROH0.5}$.

Table 3.7: Correlations (with 95% confidence intervals) between 7 different measures of autozygosity or homozygosity

	F_{ped}	F_{plink}	H_{pn}	H_{ex}	$F_{ROH0.5}$	$F_{ROH1.5}$	F_{ROH5}
F_{ped}	1	0.768 (0.711, 0.814)	0.764 (0.706, 0.811)	0.768 (0.712, 0.815)	0.844 (0.804, 0.876)	0.857 (0.820, 0.887)	0.820 (0.775, 0.857)
F_{plink}		1	0.938 (0.933, 0.942)	1.000 (1.000, 1.000)	0.744 (0.726, 0.760)	0.740 (0.722, 0.757)	0.699 (0.679, 0.718)
H_{pn}			1	0.938 (0.933, 0.942)	0.805 (0.791, 0.818)	0.796 (0.782, 0.810)	0.742 (0.724, 0.759)
H_{ex}				1	0.745 (0.728, 0.762)	0.741 (0.723, 0.758)	0.699 (0.679, 0.718)
$F_{ROH0.5}$					1	0.944 (0.940, 0.948)	0.896 (0.888, 0.903)
$F_{ROH1.5}$						1	0.949 (0.945, 0.953)
F_{ROH5}							1

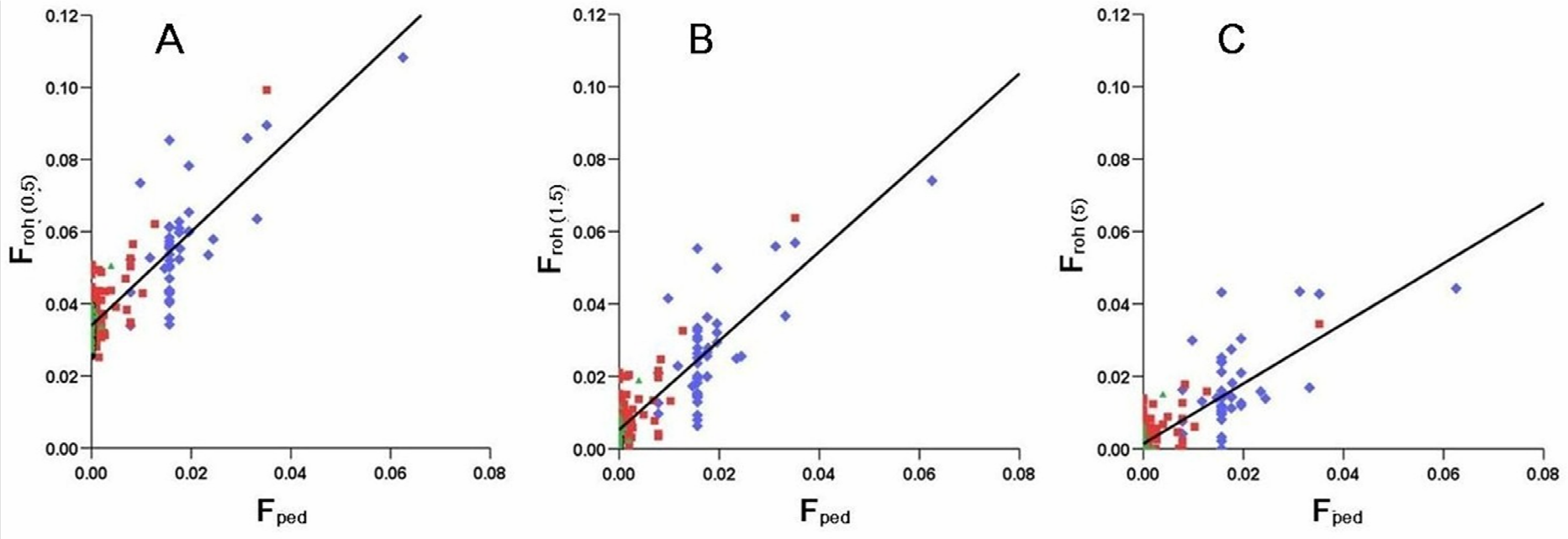
F_{ped} correlations: N = 249
All other correlations: N = 2618

A subset of the ORCADES sample (n = 249) was used to estimate correlations with F_{ped}. $F_{ROH1.5}$ was most highly correlated with F_{ped} (r = 0.857; 95% confidence interval 0.820 – 0.887). Correlations between F_{ped} and $F_{ROH1.5}$ were significantly higher than F_{ped} correlations with F_{plink}, H_{pn} or H_{ex}. The correlation between F_{plink} and F_{ped} was 0.768 (95% confidence interval 0.711 – 0.814); between H_{pn} and F_{ped} was 0.764 (0.706 – 0.811); and between H_{ex} and F_{ped} was 0.768 (0.712 – 0.815). $F_{ROH1.5}$ was slightly but not significantly more strongly correlated with F_{ped} than either $F_{ROH0.5}$ or F_{ROH5}.

Correlations between F_{ped} and $F_{ROH0.5}$, $F_{ROH1.5}$ and F_{ROH5}, are shown in figure 3.6. For each value of F_{ped} there is a range of values for F_{ROH}, reflecting stochastic variation in ancestral recombination, the existence of multiple distant parental relationships undetectable using pedigrees, and possibly pedigree misspecifications. The closer the parental relationship, the greater the absolute variance in the autozygosity of offspring. This is clear from the wide distribution of F_{ROH} values in the endogamous group compared to the mixed Orcadian group. Although as shown, ROH shorter than around 1.5 Mb do not appear to reflect differences in recent ancestral endogamy, data from the half Orcadian sample illustrate that the prevalence of these shorter ROH clearly varies between individuals. Using a minimum ROH length threshold of 5 Mb might better reflect the effects of parental relatedness on autozygosity; however it also obscures a great deal of individual genetic variation of more ancient origin. This is illustrated by the regression lines on each panel: the y-intercept gives the value of F_{ROH} where $F_{ped} = 0$. This is a measure of the proportion of the autosomes in ROH not captured by F_{ped}. Thus 0.034 of the autosomes are in

ROH longer than 0.5 Mb but are not captured by F_{ped}. The equivalent figures are 0.0053 for ROH longer than 1.5 Mb and 0.0014 for ROH longer than 5 Mb. This clearly shows that F_{ped} fails to account for autozygosity of ancient origin.

Figure 3.6: Correlations between F_{ped} and F_{ROH} in the ORCADES sample

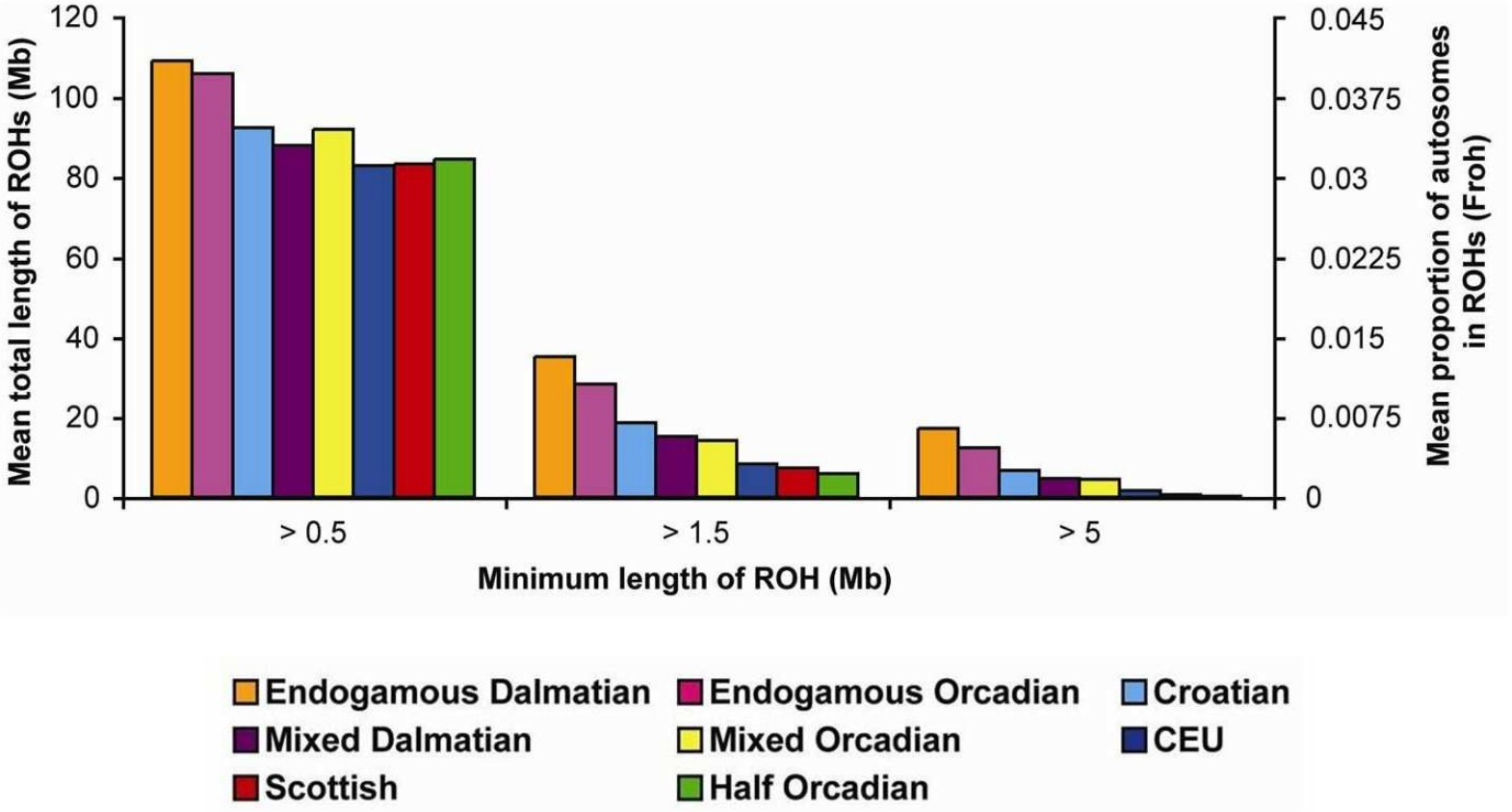


Correlations, with regression lines, are shown for 3 different minimum ROH length thresholds. Panel (a) shows the correlation between F_{ped} and $F_{ROH0.5}$; panel (b) shows the correlation between F_{ped} and $F_{ROH1.5}$ and panel (c) shows the correlation between F_{ped} and F_{ROH5}. Offspring of 1st or 2nd cousins are shown in blue; Endogamous Orcadians who are not the offspring of 1st or 2nd cousins are shown in red; Mixed Orcadians are shown in green and Half Orcadians are shown in black.

3.3.7 Mean F_{ROH} by sub-population

Mean F_{ROH} and the mean total length of ROH for each sub-population are shown in figure 3.7 for a range of minimum lengths of ROH. This figure again shows the effect on F_{ROH} in all populations of changing the ROH length cut-off point. The same 3 distinct groupings emerge for ROH longer than 1.5 Mb, although when shorter ROH are included, the picture is less clear. Using 1.5 Mb as the minimum length, endogamous Dalmatians have a mean F_{ROH} of 0.013 (35 Mb), endogamous Orcadians 0.011 (28 Mb), Croatians 0.007 (18 Mb), mixed Dalmatians 0.006 (15 Mb), mixed Orcadians 0.005 (14 Mb), CEU 0.003 (8 Mb), Scottish 0.003 (7 Mb) and half Orcadians 0.002 (6 Mb). The same relationship between groups is seen with a 5 Mb threshold, but values for all groups are reduced (to 17 Mb in endogamous Dalmatians and 0.3 Mb in half Orcadians).

Figure 3.7: Mean total length of ROH over a range of minimum ROH lengths



3.3.8 Comparison of ROH in the offspring of unrelated parents and the offspring of cousins

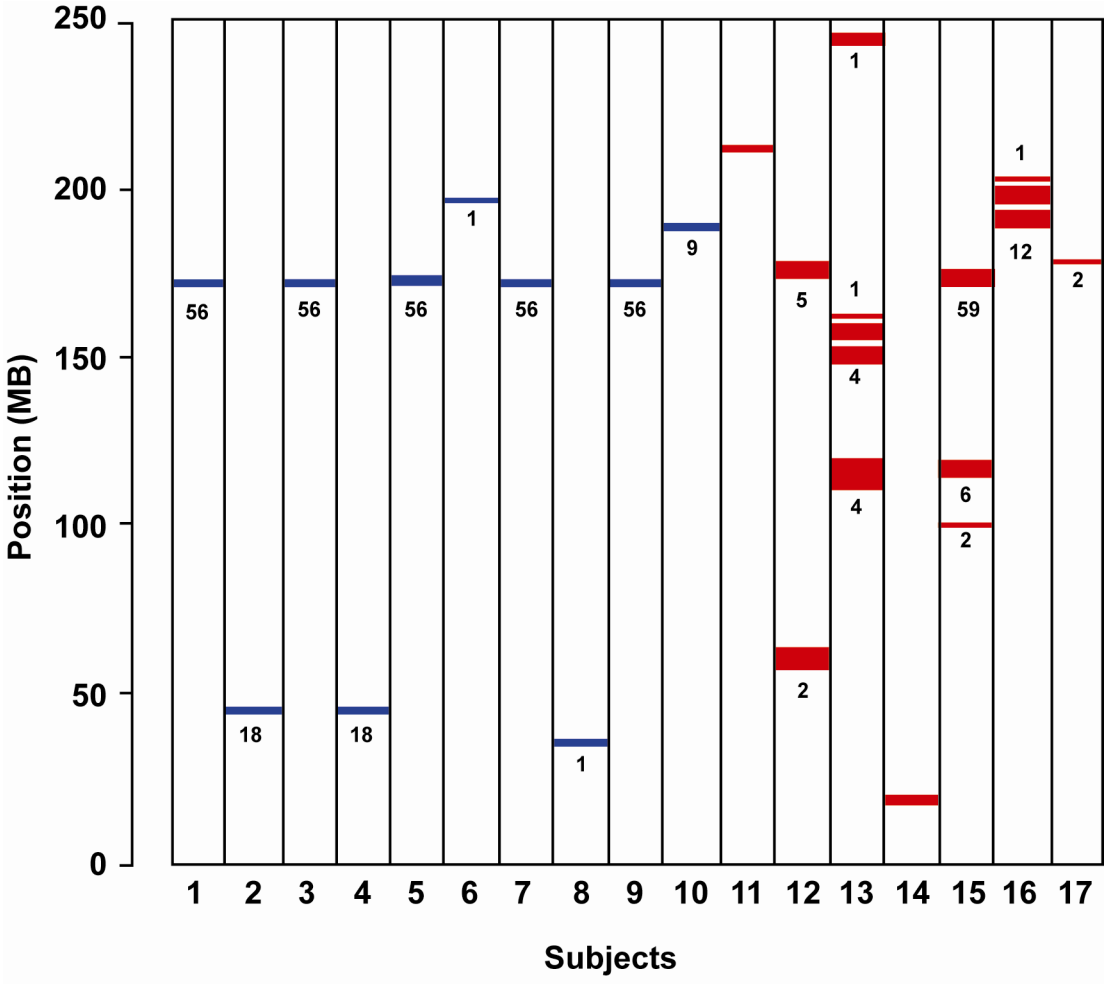
The next question for consideration was whether ROH found in half Orcadians are more common than those found in the offspring of related parents. "Common" was defined as overlapping by at least 0.5 Mb with ROH found in a subset of the Scottish sample. The number of ROH measuring ≥1.5 Mb was 143 in the half Orcadian sample, 3159 in the SOCCS control sample and 382 in the offspring of cousins sample. Results are summarised in Table 3.8. On average, each half Orcadian ROH overlapped with more than twice as many controls as did ROH in the offspring of cousins group. 12.6% of half Orcadian ROH but almost a third of ROH in the offspring of cousins group did not overlap with any of the Scottish controls. The mean number of overlaps per Mb of ROH in the 2 samples was examined in order to correct for the fact that ROH in the offspring of cousins group tend to be longer. On average, there were more than 3 times as many control overlaps per Mb of ROH in the half Orcadian group (10.9) than there were in the offspring of cousins group (3.0). Taking only those ROH measuring > 5 Mb in the offspring of cousins sample (i.e. those that are most likely to result from recent shared parental ancestry), the mean number of overlaps/Mb was only 1.4 (SD 2.0).

Table 3.8: Overlaps between ROH found in Orcadians and those found in a Scottish control sample

	Half Orcadian	Offspring of Cousins
Number of individuals	46	20
Number of ROH ≥ 1.5 Mb	143	382
Mean (SE) number of control overlaps/ROH	20.5 (22.5)	9.6 (16.0)
Maximum number of controls overlapping with a ROH	123	123
% of ROH overlapping with no controls	12.6	29
Mean (SE) number of control overlaps/Mb of ROH	10.9 (11.8)	3 (6.3)

Data on chromosome 1 for 10 individuals in the half Orcadian group (shown in blue) and 7 individuals in the offspring of cousins group (shown in red) are illustrated by way of example in Figure 3.8. These are all the individuals in the sample with ROH on chromosome 1, except that data for only one individual per sibship is shown. The numbers shown below each coloured segment are the numbers of ROH in the Scottish control sample overlapping by at least 0.5 Mb with this ROH. It is clear that although there is a tendency for ROH from both groups to cluster in certain chromosomal regions, the longer ROH in the cousin group are more randomly distributed along the chromosome.

Figure 3.8: Size and location of ROH on chromosome 1, comparing half Orcadians with the offspring of cousins



ROH measuring ≥ 1.5 Mb in 10 Half Orcadians are shown in blue and those of 7 offspring of 1st – 3rd cousins are shown in red. The numbers shown above each colored segment are the numbers of overlapping ROH in the SOCCS controls

Next all ROH in the half Orcadian group that overlapped by at least 0.5 Mb with common ROH identified by Lencz were identified (Lencz, Lambert et al. 2007). In a sample of 322 non-Hispanic European Americans, Lencz identified 339 ROH present in at least 10 subjects. 57% of the 143 half Orcadian ROH overlapped with Lencz et al's list. Only 7% (10 ROH) overlapped with neither Lencz et al's list nor the Scottish control group.

The final investigation was to determine whether the ROH in half Orcadians were found in areas of lower than average recombination. The mean recombination rate for the regions where half Orcadian ROH are located is 0.52 of the mean genome-wide recombination rate. For common ROH (i.e. half Orcadian ROH that overlap with ROH in the control group), this figure was 0.38 of the genome-wide mean.

3.4 Discussion

This study presents data on ROH for 4 populations, 2 of which are genetically isolated, with inflated levels of endogamy and inbreeding, and 2 of which are more cosmopolitan. The two isolate populations are sub-divided according to degrees of grandparental endogamy and ROH data are compared between these groups with different demographic histories in order to assess the extent to which ROH statistics reflect differences in demographic history at the population and sub-population level. An estimate of individual autozygosity (F_{ROH}) is compared with pedigree estimates of inbreeding and with 3 alternative genomic estimates of autozygosity or homozygosity to explore the utility of these approaches for estimating F in different situations. Finally, the prevalence of ROH in non-isolate populations and in individuals with demonstrably unrelated parents is quantified.

3.4.1 Copy Number Variation

In order to determine whether the ROH observed are true homozygous segments and not hemizygous deletions, CNV were analysed in the ORCADES sample. The methodology used produces a very robust estimation of the prevalence of ROH in the ORCADES sample, which to some extent over-corrects for heterozygous deletions (*personal communication*, Rehab Abdel-Rahman). Results are consistent with studies which have shown that observed ROH are true homozygous tracts and not deletions or other chromosomal abnormalities (Broman and Weber 1999; Li, Ho et al. 2006; Frazer, Ballinger et al. 2007; Simon-Sanchez, Scholz et al. 2007).

3.4.2 ROH and differences in demographic history

This study demonstrates clearly that data on ROH measuring more than 1.5 Mb accurately reflect differences in population isolation, as measured by grandparental endogamy (figures 3.1, 3.2 and 3.7). Significant differences in the mean sum of ROH were found between those with at least 3 grandparents born on the same small Orkney isle, those with at least 3 grandparents born in Orkney but not on the same isle, and those with one Orcadian and one Scottish-born set of grandparents. Significant differences were also found between those with 4 grandparents born in the same Dalmatian island village and those with 4 grandparents from the same island but born in different villages. These groups have different demographic histories in the previous 10 or so ancestral generations, although it is difficult to be precise about these timescales.

Characterising populations in terms of ROH makes it possible to situate those with unknown degrees of isolation along a spectrum. For example, beyond knowing that the SOCCS sample is broadly representative of the general Scottish population, no information is available on the precise birthplace of participants' grandparents. Data on ROH would suggest that endogamy and consanguinity are uncommon, although not unheard of in the recent ancestry of modern Scots. The 36 (4%) outliers in the Scottish sample with ROH suggestive of parental relatedness (total ROH $\geq$ 5 Mb) were no more likely to live in rural or island locations than in urban locations. This is unsurprising: Scotland is a small, largely urbanized country with high population mobility and considerable immigration. There are, however, small, remote island communities off the west and north coasts of Scotland which have been shown to

have greater LD and lower haplotype diversity than mainland urban and rural Scottish populations (Vitart, Carothers et al. 2005), consistent with lower effective population sizes, isolation and genetic drift. Orkney is one such isolated community, however as is shown here, even within such small populations, there is a great diversity of ancestry, from the tightly endogamous to the completely outbred. These data show that having at least 3 grandparents from within a 2-3 mile radius (as is the case in the North Isles of Orkney and Dalmatian villages) is associated with considerably more and longer ROH than merely coming from Orkney or a Dalmatian island.

The distribution of ROH in the CEU sample, which is widely used as a northwest European reference population, does indeed appear to be very similar in this respect to that in the Scottish sample. Consistent with other studies (Gibson, Morton et al. 2006; Frazer, Ballinger et al. 2007), this approach picks out one outlier, who is likely to be the offspring of consanguineous parents (NA12874, or CEPH1459-11).

The Dalmatian sub-sample of offspring of Croatian settlers is similar to the mixed Dalmatian and mixed Orcadian subgroups by various ROH-based measures, suggesting that these settlers came from fairly small, semi-isolated communities where endogamy was not uncommon.

3.4.3 ROH and inbreeding

The hypothesis under investigation here is that the proportion of the autosomal genome in ROH (F_{ROH}) provides a reliable estimate of the effects of parental

relatedness because autozygous genotypes are not evenly distributed throughout the genome but are distributed in runs or tracts (Figure 1.1). Thus extended tracts of homozygosity are a signature of inbreeding and this can be quantified by measuring them and expressing them as a proportion of the typed autosomal genome.

The hypothesis was tested by comparing F_{ROH} statistics with F_{ped} statistics derived from high quality pedigree data complete to 5 ancestral generations in the ORCADES sample. The correlation between F_{ROH} and F_{ped} was then compared with correlations between F_{ped} and 3 other genomic estimates of autozygosity or homozygosity (see chapter 1 section 2) to assess which approach best estimates the genomic effects of recent parental relatedness.

This study shows that F_{ROH} is strongly correlated with F_{ped} derived on the basis of 5 ancestral generations, and significantly more so than the other 3 measures investigated (H_{pn}, H_{ex}, F_{plink}). Perfect correlation is not expected, largely because of the deficiencies of F_{ped}, which cannot account for stochastic variation, pedigree inaccuracies or the effects of multiple inbreeding loops just beyond the limits of the available ancestral information. This is particularly the case in the ORCADES sample, where multiple distant parental relationships 6, 7 and 8 ancestral generations in the past but undetectable with only 5 ancestral generations of pedigree information, inflate autozygosity, such that the offspring of these distant cousins can be almost as autozygous as the offspring of first cousins (Liu, Elefante et al. 2006). The individual with the second highest F_{ROH} in the ORCADES sample, for example, is the offspring of a couple whose closest relationship is that of 3rd cousins but who

are multiply related at least 24 different ways in the last 8 generations alone (red outlier, figure 3.6). Whilst $F_{ROH0.5}$ for this individual is almost as high as that for the highest individual in the sample, who is the offspring of first cousins, F_{ped} is almost half the value of F_{ped} for the first cousin offspring. Thus F_{ped} fails to capture all of the autozygosity estimated by F_{ROH}.

The y-intercepts of the regression lines in the 3 panels of figure 3.6 quantify the under-estimation of autozygosity by F_{ped} compared with F_{ROH}. In general, the shorter the ROH, the more distant the common maternal and paternal ancestor from whom it originated. This means that the lower the minimum length of ROH used in F_{ROH} estimation, the further back in time it is possible to look and the more distant the parental relationship it is possible to detect. This is clearly illustrated by the regression lines in figure 3.6: using a 5 Mb cut-off, there is very little difference between the amount of autozygosity estimated by F_{ped} on the basis of 5 ancestral generation pedigrees, and that detected by F_{ROH}. This suggests that in this population, ROH of 5 Mb or longer are sufficient to detect inbreeding resulting from 3rd cousin parental relationships or closer. Decreasing the minimum ROH length threshold to 1.5 or 0.5 Mb increases the gap between F_{ROH} and F_{ped} estimates, which illustrates that using a lower minimum ROH length cut-off allows detection of the effects of more distant inbreeding, beyond the range of available pedigree data. The fact that $F_{ROH1.5}$ statistics accurately reflect the effects of multiple distant inbreeding in one individual known to have multiple inbreeding loops 6, 7, 8 and more generations in the past suggests that ROH of this length are sufficient to detect the effects of inbreeding beyond the range of available pedigree data (although it is

difficult to be any more precise about the relationship between ROH length and pedigree depth because of inherent stochastic variation).

3.4.4 ROH in outbred subjects

ROH can, then, be used to estimate the effects of recent parental inbreeding at the individual level. F_{ROH} can be used to assess inbreeding effects on disease or QT. Using a homozygosity mapping approach, specific ROH can also potentially be used as a means of narrowing down the search for variants influencing disease or QT. Not all ROH, however, are indicative of parental relatedness. The Phase II HapMap study estimates that ROH measuring in excess of around 100 kb constitute 13-14% of the genome in Europeans (Frazer, Ballinger et al. 2007). Lencz et al. (Lencz, Lambert et al. 2007) give a similar estimate. The findings of the present study are not directly comparable, as ROH shorter than 500 kb are not examined here; however it is shown (figure 3.1) that ROH measuring between 500 and 1500 kb were present in all individuals in all the sub-populations studied, totaling on average 75 Mb per individual (2-3% of the autosomal genome). The fact that small but significant differences were found *among* the 4 study populations in the mean sum of these short ROH, but no significant differences were found *within* populations (e.g. between endogamous Orcadians and half Orcadians), lends support to the view that population differences in the prevalence of ROH shorter than around 1.5 Mb reflect LD patterns of ancient origin, rather than the effects of more recent endogamy. Short ROH are, then, ubiquitous in the genomes of outbred individuals but it is not unusual to find longer ROH in their genomes also. This study shows that ROH measuring up to several Mb in length are not uncommon in demonstrably outbred

individuals (the half Orcadian sample, who are known to have no common maternal and paternal ancestor in 5 and probably at least 10 generations) and in populations where inbreeding is rare (SOCCS and CEU). These findings are consistent with a number of recent observational studies using high density genome scan data, which have suggested that ROH longer than 1 Mb are more common than previously thought in outbred individuals (Broman and Weber 1999; Gibson, Morton et al. 2006; Li, Ho et al. 2006; Lencz, Lambert et al. 2007; Simon-Sanchez, Scholz et al. 2007; Curtis, Vine et al. 2008).

The picture of genome-wide homozygosity now emerging is that short stretches measuring tens of kb and indicative of ancient LD patterns are common, covering up to one third of the autosomal genome (Frazer, Ballinger et al. 2007). At the other end of the spectrum, very long ROH, measuring tens of Mb, are the signature of parental relatedness. In between, ROH may result from recent parental relatedness or may be autozygous segments of much older pedigree, which have occurred because of the chance inheritance through both parents of extended haplotypes that are at a high frequency in the general population, possibly because they convey or conveyed some selective advantage (Lencz, Lambert et al. 2007) or possibly simply because of genetic drift. Where such haplotypes are located in regions of the genome where recombination rates are low, they may extend to several Mb in length. Other studies have suggested that ROH cluster in such low recombination genomic regions (Gibson, Morton et al. 2006; Li, Ho et al. 2006; Simon-Sanchez, Scholz et al. 2007; Curtis, Vine et al. 2008) and the data from the present study support this.

The fact that ROH, even quite long ROH, appear to be common in outbred subjects suggests two strands for further investigation. Firstly, if the aim is to investigate recessive effects, accurate quantification at the individual level of the shorter of these ROH is important. This presents methodological challenges, which will be explored further below and in the following chapter. Secondly, the growing realization that ROH are common in outbred individuals has sparked interest in identifying and investigating these as specific disease risk factors. Consistent with the findings of other studies (Lencz, Lambert et al. 2007; Curtis, Vine et al. 2008), the present study shows that ROH in outbred subjects are almost invariably common (i.e. shared or overlapping between several individuals) but not universal. The ORCADES sample had ROH overlapping with those occurring in both the SOCCS sample (table 3.8 and figure 3.8) and in an outbred non-Hispanic European American population (Lencz, Lambert et al. 2007). Common ROH are, then, a source of individual genetic variation which may play a causal role in common complex disease, in as far as (partial) recessive effects are found in these diseases, and which therefore merit further exploration as risk factors in their own right (Lencz, Lambert et al. 2007).

3.4.5 Using genomic measures to estimate homozygosity, autozygosity and inbreeding

This chapter aimed to assess the utility of F_{ROH} as a measure of individual autozygosity and to compare F_{ROH} with 3 other genomic approaches to quantifying autozygosity or homozygosity. This turns out not to be quite as straightforward as it might initially have appeared. Page 1 of this thesis contrasts autozygosity, where identical alleles are inherited from a common ancestor, with “chance homozygosity”,

where the alleles are identical by state but not by descent. However, the results of this and other recent observational studies present a direct challenge to this established way of thinking about homozygosity and autozygosity. If it were possible to look directly at the genome, as will become increasingly the norm with the 1000 Genomes project (www.1000genomes.org (Durbin and Altshuler 2008)), what would be apparent would be a pattern of homozygous segments of different lengths: some very short, originating from very ancient shared ancestry; and some longer, possibly resulting from more recent shared parental ancestry or possibly ancient haplotypes located in low recombination genomic areas. Regardless of their history, however, these homozygous segments have been shaped by exactly the same forces: all ROH are the result of the chance inheritance from both parents of identical chromosomal segments. As such, all ROH are autozygous: all are inherited from common maternal and paternal ancestors, albeit possibly very distant. By the same token, the distinction between autozygosity and chance homozygosity is a false one: on the one hand, all homozygosity can be said to occur by chance, because meiosis is a highly random process; on the other, an apparently sporadic homozygous genotype is no more the result of chance than a homozygous genotype located in a ROH. Even the commonly used notion of sporadic homozygosity is ill-conceived: the appearance of being sporadic is highly dependent on the density of SNP panel being used to make the observation. With a denser SNP panel, what appeared to be a lone homozygous genotype may be seen to be part of a very short ROH. According to this way of thinking, there is no inherent distinction between homozygous and autozygous genotypes, just a difference in the degree of parental relatedness (although this assertion has to be qualified somewhat: very short ROH resulting

from shared parental ancestry deep in the past are more likely than longer ROH of more recent origin to harbour rare, unobserved heterozygous genotypes because there has been more time for mutation).

It follows from this that the ideal way to quantify total individual homozygosity would be to identify and sum all ROH, from the shortest to the longest. Defining a chromosomal segment as a ROH is essentially a method of using SNP data to infer the homozygosity status of the intervening, unobserved stretches of the chromosome. The aim is to maximize the probability that between the first and last observed SNPs in the ROH, the entire unobserved stretch of the chromosome is homozygous. The length of ROH that can be detected by this approach is, therefore, highly sensitive to the density of SNP panel being used. For example, if a chromosomal segment 100 kb long contains only 5 SNPs, this would not be a reliable basis for predicting the homozygosity status of the segment as a whole. Clearly, the shorter the ROH, the denser the scan needed to detect it reliably. This is considered in more detail in the next chapter, where it is shown that, whilst a panel of 300,000 SNPs detects ROH longer than 1.5 Mb with a high degree of reliability, estimates become increasingly unreliable the shorter the ROH below this level.

Whilst this study has shown that F_{ROH} derived from a 300,000 SNP panel is a reliable measure of recent parental relatedness which can therefore be used to investigate inbreeding effects in isolate or consanguineous populations, this may not be the most suitable approach to quantifying individual homozygosity in order to quantify

recessive effects in more cosmopolitan populations. Denser SNP panels or other approaches may be more fruitful.

All 3 alternative measures considered here (H_{pn}, F_{plink} and H_{ex}) to varying degrees estimate not just homozygosity resulting from recent inbreeding but also homozygosity resulting from much more ancient shared parental ancestry. Thus in contrast to F_{ROH}, these 3 measures do not pick up significant differences between half and mixed Orcadians (figures 3.3 and 3.4 and table 3.4) – i.e. differences in degrees of parental relatedness originating within the past 8-10 generations. On the other hand, mean H_{pn}, F_{plink} and H_{ex} estimates of SOCCS subjects with 4 Scottish-born grandparents were significantly higher than the equivalent estimates of those with at least one grandparent born outside Scotland; whereas there was no significant difference between these 2 groups in mean F_{ROH}. In other words, these 3 alternative measures can detect the effects of very distant parental relatedness more effectively than is possible with ROH derived from a 300,000 SNP panel and as such may be more suitable than ROH-based measures for investigating recessive effects in non-isolate general populations.

Chapter 4: Measuring short ROH

4.1 Introduction

The original objective of this study was to explore whether genomic data on ROH could be used to investigate recessive effects in an inbred sample. What has emerged from the analysis described in chapter 3 is that demonstrably outbred individuals differ from one another in the number, length and location of ROH in their genomes and in their levels of homozygosity as estimated by a variety of measures, and that these are aspects of individual genetic variation which merit investigation as potential disease risk factors. The purpose of this chapter is to investigate the utility of F_{ROH} for quantifying homozygosity in outbred samples.

A homozygous genotype originates in one of two ways: either (rarely) through mutation of an allele in what would otherwise have been a heterozygous genotype, or (commonly) through the inheritance from both parents of an identical ancestral allele. If the common maternal and paternal ancestor from whom the allele originates is a fairly recent one, it will tend to be inherited as part of a long sequence of homozygous genotypes or ROH. Such long ROH consist of identical copies of chromosomal segments inherited through both parents which have not been broken down into small segments by repeated meioses because they originate from an ancestor only a few generations in the past. The analysis of the ORCADES sample presented in chapter 3 demonstrates that F_{ROH} derived using a 300,000 SNP panel and a minimum ROH length of 1.5 Mb correlates strongly with F_{ped} and therefore provides a reliable estimate of parental relatedness originating in the previous 5 – 10

ancestral generations. As such, F_{ROH} provides a useful approach to investigating inbreeding effects in isolated or consanguineous populations.

The more distant the common ancestor, the shorter the sequence of homozygous genotypes, or ROH, is likely to be (although this is not always the case: ROH longer than 1 or 2 Mb, and typically found in genomic regions where recombination rates are low, are not uncommon in demonstrably outbred individuals). Shorter ROH (up to 1 or 2 Mb in length) are extremely abundant throughout the genome (Frazer, Ballinger et al. 2007). These shorter ROH are produced by exactly the same mechanism as the longer ROH that are generally indicative of inbreeding: identical haplotypes originating from a common ancestor, typically many, many generations in the past, are inherited through both parents. These haplotypes are often very common: they may have reached high frequency in the population because in the past they conferred some selective advantage or perhaps as a result of a special form of genetic drift called allelic surfing (Hofer, Ray et al. 2009).

The only truly sporadic homozygotes are those that have arisen through mutation. Because the vast majority of homozygous genotypes arise through the inheritance from both parents of an identical ancestral allele, albeit generally originating many, many generations ago, it follows that homozygous genotypes are typically found in ROH and that if it were feasible routinely to sequence the genomes of study subjects, quantifying individual autozygosity would be a simple matter of identifying and summing each subject's ROH, down to the very shortest ones. However in reality, at present, ROH can only be identified using panels of 300,000, 500,000 or 1 million

SNPs. This may change in the near future, when the 1000 Genomes Project data become available (Durbin and Altshuler 2008) and it becomes possible to impute a much larger number of variants to enhance the utility of existing SNP panels; however with existing technology there is a limit on the length of ROH it is possible to detect reliably. The high correlation between F_{ped} and F_{ROH} described in chapter 3 suggests that the Illumina Hap300 genotyping platform can accurately detect the longer ROH that are indicative of inbreeding; however the comparison outlined in chapter 3 between those with 4 and those with fewer than 4 Scottish-born grandparents in the SOCCS sample suggests that H_{pn} or other approaches might provide a more reliable means of detecting the effects of more ancient parental relatedness. Estimates of H_{pn} were significantly higher in those with 4 Scottish-born grandparents than in those with fewer than 4 Scottish-born grandparents. Those with 4 Scottish-born grandparents share more maternal and paternal ancestry (albeit very distant) than those with fewer than 4 Scottish-born grandparents. In other words, there are more (distant) inbreeding loops in their pedigrees than are present in the pedigrees of those with at least one non-Scottish-born grandparent. Referring back to the underlying biological mechanisms (figure 1.1), the expectation is that those with 4 Scottish-born grandparents will have more of their autosomal genome in short ROH than will those in the comparison group and this is reflected in estimates of H_{pn}. In the absence of denser SNP panels, and assuming that long and short ROH exert similar influences, H_{pn} might, then, be a better measure of homozygosity than F_{ROH} for investigating recessive effects in outbred samples. This chapter explores this issue by seeking to answer two questions:

- What is the minimum length of ROH that can be detected using a 500,000, a 300,000 and a 50,000 SNP panel? The densest currently available SNP dataset is the HapMap release 23a, which contains over 3 million SNPs for 270 individuals in 4 populations. Beyond complete sequencing of subjects' genomes, this panel is as close as it is currently possible to get to direct observation. Using the Utah European American (CEU) data in the 60 founders (CEU parents, corresponding to CEPH grandparents) for this panel as a baseline, this study compares the 3 different densities of SNP panel to assess the minimum length of ROH each panel can reliably detect.
- What proportion of observed homozygous genotypes is in ROH of different length categories? In particular, what proportion is in ROH shorter than can be reliably detected using a 300K or a 500K SNP panel? This gives an indication of the proportion of total homozygosity not captured by F_{ROH} statistics – in other words, the proportion of homozygosity resulting from an individual's very distant ancestry.

4.2 Methods

4.2.1 CEU sample details

HapMap release 23a (termed here the 2400K SNP panel) for CEU founders (n = 60) was downloaded from the HapMap website (www.hapmap.org (HapMap 2002)). Although this was the densest available SNP panel at the time of the analysis (December 2008) it is important to recognise that it is not sequence-based but is itself a snapshot of genomic variation which samples only a subset of the predicted total number of SNPs. SNPs with a minor allele frequency of < 1%, SNPs with > 10%

missing genotypes and SNPs failing the Hardy Weinberg Equilibrium test at $p < 0.0001$ were removed. All individuals in the sample had fewer than 5% missing genotypes, so all were retained. Three further SNP panels, with approximately 500,000, 300,000 and 50,000 SNPs, were then derived from this to reflect the data available on cohorts with health information, as summarised in table 4.1. The 500K panel consists of SNPs present in both the cleaned CEU release 23a panel and a panel consisting of the combined Illumina HumanHap300 and Illumina HumanHap240S chips. Similarly, the 300K panel used here was derived from the overlap between the Illumina Infinium HumanHap300 v2 platform, the combined Illumina HumanHap300 and HumanHap240S chips and the cleaned CEU release 23a panel. The fourth (50K) panel was derived by pruning the resulting 300K panel for linkage disequilibrium (LD) using the *pairwise* routine in PLINK (Purcell, Neale et al. 2007). A window of 100 SNPs was moved across the genome, in steps of 25 SNPs. One of each pair of SNPs in strong LD with each other ($r^2 > 0.2$) was removed. This panel was included because the PLINK website recommends pruning the panel for strong LD if the aim is to identify autozygous segments as opposed to what it terms ROH that are "homozygous by chance" (Purcell 2007). The first and last SNP for each chromosome and the SNPs before and after each centromere in the LD-pruned SNP panel (termed here the 50K SNP panel) were also used as the boundaries for the other SNP panels to ensure like-for-like comparison: the only difference between the four SNP panels used in the comparison is therefore in the density of SNP coverage.

Table 4.1: Description of SNP panels

Infile	N snps	Outfile	N snps	Details
CEU founders release 23a	3,849,032	Cleaned CEU founders release 23a	2,422,661	Removal of SNPs with: • MAF < 0.01 (1,319,280 SNPs failed), • > 10% missing (138,266 failed). No SNPs failed HWE test at p<0.0001 and no individuals failed missingness test at >5% missing genotypes.
Cleaned CEU founders release 23a	2,422,661	Overlap between: • Cleaned CEU founders release 23a • Combined Illumina HumanHap300 and Illumina HumanHap240S	495,426	Combined Illumina HumanHap300 and Illumina HumanHap240S has 534,506 autosomal SNPs. 39,080 SNPs are present in combined Illumina panels but not in cleaned CEU founders release 23a.
Overlap between: • Cleaned CEU founders release 23a • Combined Illumina HumanHap300 and Illumina HumanHap240S	495,426	Overlap between: • Cleaned CEU founders release 23a • Combined Illumina HumanHap300 and Illumina HumanHap240S • Illumina Infinium HumanHap300 v2	302,703	Illumina Infinium HumanHap300 v2 has 309,200 autosomal SNPs. 6,497 SNPs are present in Illumina Infinium HumanHap300 v2 but not in cleaned CEU founders release 23a overlap with combined Illumina HumanHap300 and Illumina HumanHap240S .
Overlap between: • Cleaned CEU founders release 23a • Combined Illumina HumanHap300 and Illumina HumanHap240S • Illumina Infinium HumanHap300 v2	302,703	50K panel	52,888	This file was derived by taking the CEU founders release 23a, combined Illumina HumanHap300 and Illumina HumanHap240S and Illumina Infinium HumanHap300 v2 overlap file and using the pairwise LD pruning option in PLINK to remove one of each pair of SNPs in strong LD (r^2 > 0.2). Panel length: 2657.85 Mb Density of SNP coverage: 50.25 kb/SNP

Table 4.1 continued: Description of SNP panels

Infile	N snps	Outfile	N snps	Details
Overlap between: • Cleaned CEU founders release 23a • Combined Illumina HumanHap300 and Illumina HumanHap240S • Illumina Infinium HumanHap300 v2	302,703	300K panel	301,944	This file was derived by making the same chromosomal and centromeric boundaries as the 50K panel. Panel length: 2657.85 Mb Density of SNP coverage: 8.8 kb/SNP
Overlap between: • Cleaned CEU founders release 23a • Combined Illumina HumanHap300 and Illumina HumanHap240S	495,426	500K panel	494,203	This file was derived by making the same chromosomal and centromeric boundaries as the 50K panel. Panel length: 2657.85 Mb Density of SNP coverage: 5.38 kb/SNP
Cleaned CEU founders release 23a	2,422,661	2400K panel	2,412,807	This file was derived by making the same chromosomal and centromeric boundaries as the 50K panel. Panel length: 2657.85 Mb Density of SNP coverage: 1.1 kb/SNP

4.2.2 Definition and analysis of ROH

Data were analysed using the Runs of Homozygosity program implemented in PLINK (Purcell, Neale et al. 2007), as described in chapter 3. The minimum number of consecutive homozygous SNPs constituting a ROH was set at 25; the minimum length of ROH was set at 150 kb and the maximum gap allowed between consecutive SNPs in a ROH was set at 100 kb. The minimum density of a ROH was set at 50 kb/SNP for the 2400K, 500K and 300K panels but no density criterion was specified for the 50K panel, as SNP-coverage was so sparse in this panel that specifying a density threshold would have rendered the panel unusable.

Data were then transferred into SPSS to estimate F_{ROH} statistics as described in the previous chapter. For ease of interpretation, F_{ROH} statistics are expressed here as percentages rather than proportions, with the subscript number referring to the minimum length of ROH (Mb) included in the statistic. Thus $F_{ROH1.5}$ is defined as the percentage of the typed autosomal genome in ROH longer than 1.5 Mb.

4.2.3 Sensitivity analysis

Correlations of the results between the 2400K SNP panel and each of the 3 other panels were calculated for each length category and threshold in order to assess the reliability of using the 3 panels to measure ROH of different lengths.

In chapter 3, ROH data from the CEU sample were compared with data from an Orcadian sample with very reliable pedigree data. Whilst the majority of the CEU

sample used here was very similar to an outbred sub-group of the Orcadian sample, with no known parental relatedness in at least 5 and probably at least 10 ancestral generations (mean $F_{ROH1.5}$ = 0.24%, as measured with the Hap300 panel), one individual (NA12874; CEPH 1459-11) was found to be similar to the sub-group of individuals who are the offspring of first or second cousins ($F_{ROH1.5}$ = 3.51%, measured in the same way). CEPH1459-11 is the maternal grandfather of the family and no pedigree data are available to confirm this; however these findings are consistent with Gibson et al (2006), who also identified this individual as being the probable offspring of related parents because long ROH were located in genomic regions with low levels of LD. Because the presence of this outlier may skew results, data were therefore analysed both with and without this individual.

4.2.4 Proportion of observed homozygous genotypes in ROH

Observed homozygous calls for each individual were counted using the 2400K panel in PLINK. The number of SNPs in ROH longer than 150 kb, 300 kb, 600 kb, 1200 kb and 2400 kb per individual was expressed as a proportion of the total number of observed homozygous genotypes for each individual. Sample means were then estimated and graphed in order to estimate the proportion of observed homozygous genotypes outside ROH of various lengths.

4.3 Results

4.3.1 F_{ROH} estimates using SNP panels of different densities

Population distributions of the proportion (%) of the typed autosomal genome in ROH (i.e. F_{ROH} expressed as a percentage) are shown in Figure 4.1 for each panel

and for a range of minimum ROH length cut-offs. Compared with the 2400K SNP panel all panels underestimate the percentage of the typed autosomal genome in ROH longer than 0.15, 0.35 and 0.5 Mb (Figure 4.1 a-c), and this underestimation is more marked the less dense the panel and the smaller the minimum ROH length cut-off. These data are also shown in table 4.2. Using the 2400K panel, on average around 18% of the autosomal genome is in ROH longer than 150 kb. This compares with around 9% using the 500K panel, 4% using the 300K panel and 0.02% with the 50K panel.

Figure 4.2 is a scatter plot showing the proportion (%) of the typed autosomal genome in ROH longer than 150 kb using the 2400K and 500K panels. A line of best fit is shown, which intercepts the y-axis at 11.1. This means that the 2400K panel estimates that 11.1% more of the typed autosomal genome is in ROH longer than 150 kb compared with the 500K panel estimate.

In contrast, the higher the minimum ROH threshold, the closer the agreement between the different panels. There is no significant difference between the 2400K panel and the 500K panel in estimates of mean F_{ROH1} (figure 4.1 d – f), and this is the case regardless of whether or not the outlier is included. The estimate of mean F_{ROH1} for the 2400K panel (outlier included) is 1.09% (95% confidence interval 0.99-1.18). The estimate of mean F_{ROH1} for the 500K panel is 1.07% (95% confidence interval 0.95 – 1.18). Mean F_{ROH1} excluding the outlier is 1.05% (0.98 to 1.11) for the 2400K panel and 1.01% (0.96 – 1.07) for the 500K panel. Figure 4.3 shows $F_{ROH1.5}$ using the 2400K and 500K panels. A line of best fit is shown, which intercepts the y-axis

at 0.1. This means that the 2400K panel estimates that only 0.1% more of the typed autosomal genome is in ROH longer than 1500 kb compared with the 500K panel estimate.

Comparing the 300K panel with the 2400K panel, there is no significant difference between the two panels for F_{ROH1} when the outlier is included (0.94%, 95% confidence interval 0.82 – 1.05 for the 300K panel); although when the outlier is excluded, the 300K panel estimate of F_{ROH1} is significantly lower than the 2400K panel estimate (0.88%, 95% confidence interval 0.83 – 0.94). Excluding the outlier, there is no significant difference between the 2400K and 300K panel estimates for $F_{ROH1.5}$ (0.35%, 95% confidence interval 0.27 to 0.42 using the 2400K panel; 0.31%, 95% confidence interval 0.19 to 0.42% using the 300K panel).

Very few ROH were detected using the 50K panel, none longer than 1.5 Mb, and this was the only panel which failed to identify the outlier.

Differentiation is poor in all four panels for ROH longer than 2 Mb because ROH of this length are uncommon in this outbred sample (median number of ROH > 2 Mb per person = 2 for 2400K panel, 1 for 500K and 300K panels and 0 for 50K panel; mean number of ROH > 2 Mb per person = 1.6 for 2400K panel, 1.5 for 500K panel, 1.3 for 300K panel and 0 for 50K panel).

Figure 4.1: Population distributions for the proportion (%) of the typed autosomal genome in ROH for a range of minimum ROH length cut-offs, using each of the 4 SNP panels

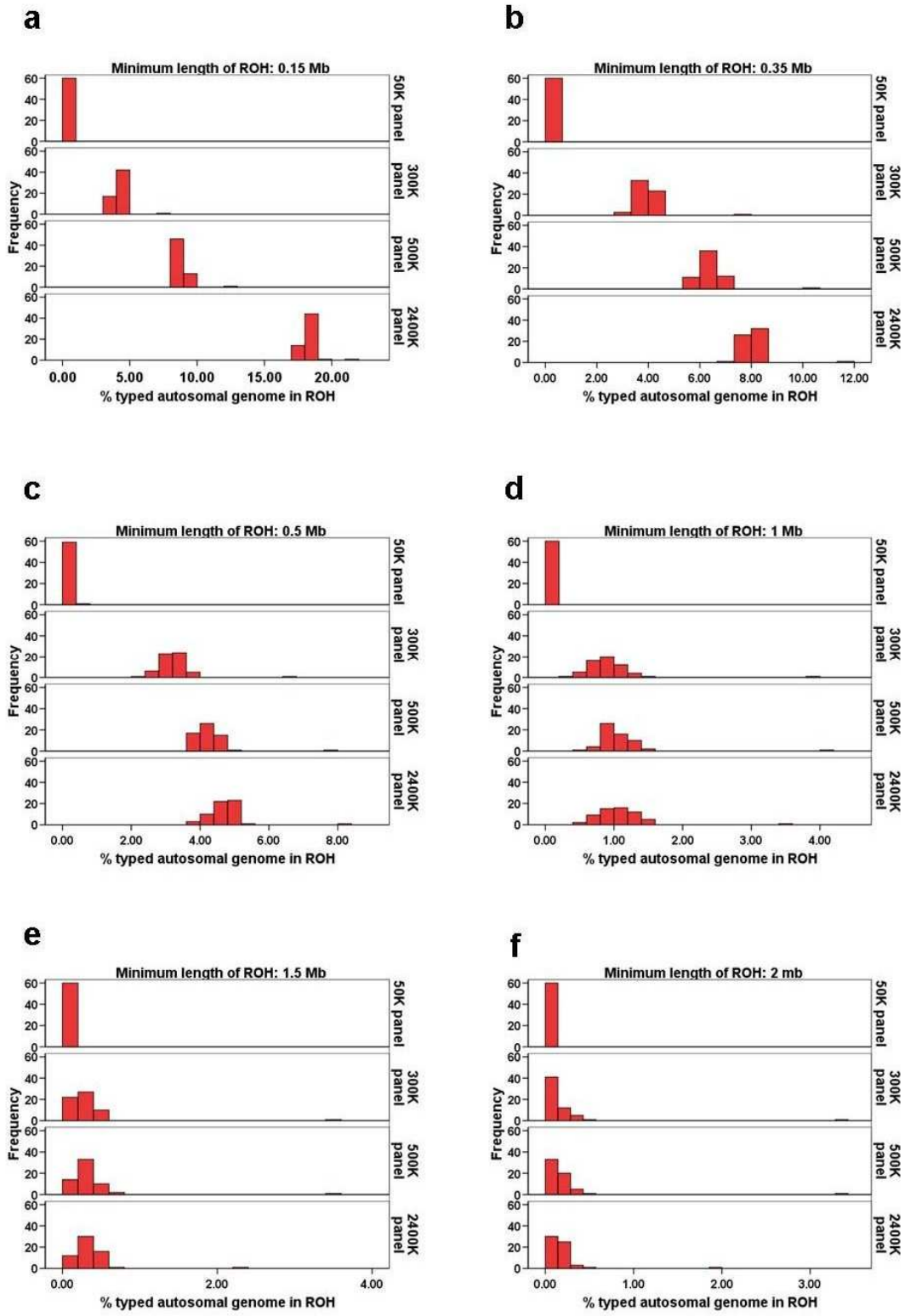


Table 4.2: Mean proportion (%) of the typed autosomal genome in ROH by SNP panel and minimum ROH length cut-off, including and exluding the outlier

Minimum length of ROH (kb)	Outlier	Mean (SE) % autosomal genome in ROH			
		2400K	500K	300K	50K
150	incl	18.37 (0.072)	8.78 (0.078)	4.26 (0.074)	0.016 (0.01)
	excl	18.31 (0.048)	8.72 (0.049)	4.2 (0.044)	0.006 (0.003)
350	incl	8.06 (0.075)	6.38 (0.078)	3.96 (0.075)	0.016 (0.01)
	excl	7.99 (0.044)	6.31 (0.047)	3.9 (0.044)	0.006 (0.003)
500	incl	4.7 (0.069)	4.29 (0.073)	3.2 (0.072)	0.014 (0.009)
	excl	4.65 (0.041)	4.22 (0.038)	3.14 (0.041)	0.005 (0.003)
1000	incl	1.09 (0.05)	1.07 (0.06)	0.94 (0.06)	0.005 (0.003)
	excl	1.05 (0.031)	1.01 (0.027)	0.88 (0.03)	0.002 (0.002)
1500	incl	0.35 (0.038)	0.36 (0.055)	0.31 (0.058)	
	excl	0.31 (0.018)	0.31 (0.018)	0.25 (0.018)	
2000	incl	0.16 (0.033)	0.18 (0.056)	0.16 (0.057)	
	excl	0.13 (0.014)	0.13 (0.015)	0.1 (0.014)	

Figure 4.2: Estimated proportion (%) of the typed autosomal genome in ROH longer than 150kb using 2400K and 500K panels

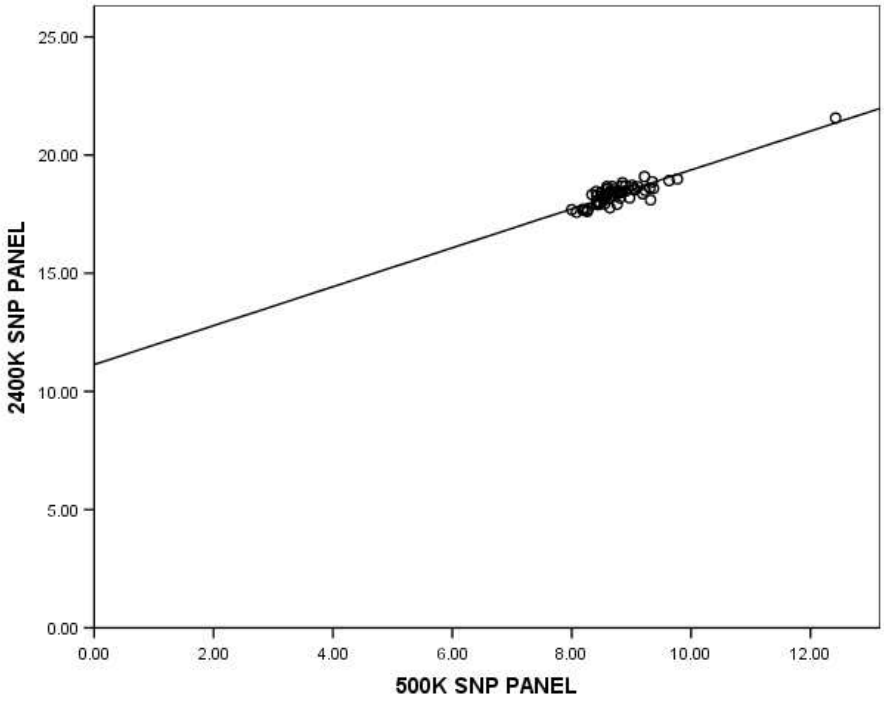
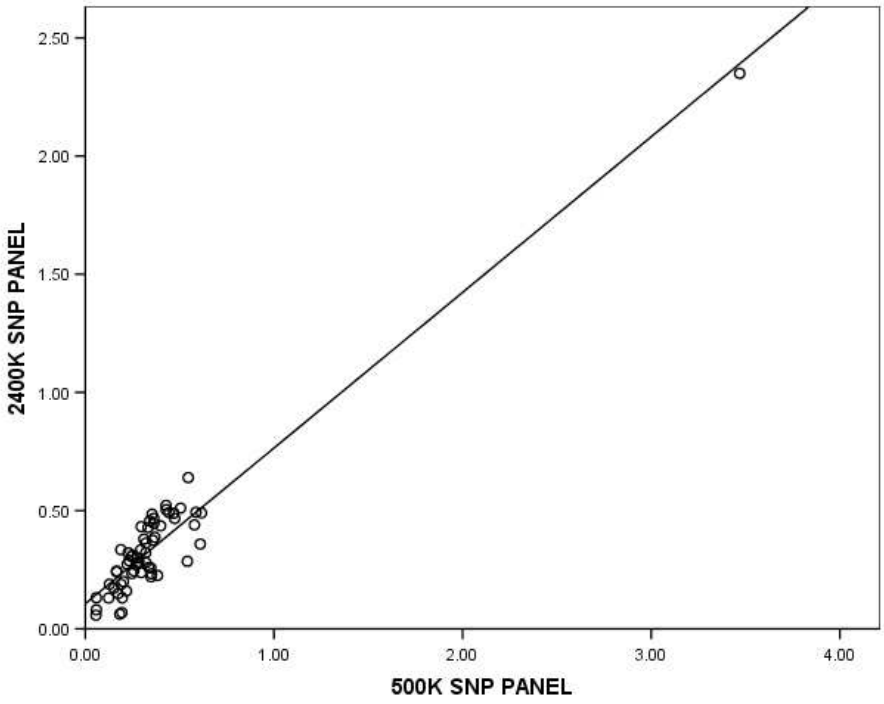


Figure 4.3: Estimated proportion (%) of the typed autosomal genome in ROH longer than 1500kb using 2400K and 500K panels



4.3.2 Correlation between SNP panels

Correlation coefficients between the 2400K SNP panel and each of the other panels are shown graphically in Figures 4.4 and 4.5 and in tables 4.3 and 4.4. Firstly, F_{ROH} correlations were estimated, using a range of minimum ROH length thresholds (figure 4.4). All panels correlate more strongly with the 2400K SNP panel with the inclusion of the outlier than with the outlier excluded, suggesting that that there is closer agreement with this dense panel for longer rather than shorter ROH ($r = 0.9$ for the 500K panel including the outlier, compared with 0.7 – 0.8 excluding the outlier; $r \sim 0.85$ for the 300K panel with the outlier compared with ~ 0.7 without; $r \sim 0.7$ for the 50K panel with the outlier but the correlation coefficient is not significantly different from zero without the outlier). The same information, but compressed into ROH longer and shorter than 1 Mb, is summarised in figure 4.5. The 500K and 300K panels correlate moderately ($r \sim 0.4$ - 0.6) with the 2400K SNP panel for ROH shorter than 1 Mb and strongly ($r \sim 0.8$ - 0.9) for ROH longer than 1 Mb. The only significant correlation found between the densest panel and the 50K panel was for ROH longer than 1 Mb with the inclusion of the outlier ($r \sim 0.8$). Otherwise, the correlation coefficient was not significantly different from zero. This illustrates that the 50K panel, which was derived by stripping out SNPs in strong LD, will by definition be unable to detect ROH arising from short haplotypes inherited as blocks. This panel is also not dense enough to detect longer ROH which are not simply reflective of LD: as illustrated in figure 4.1, the 50K panel did not detect the outlier.

When data are analysed by length category rather than a minimum length threshold, correlations are generally weaker (table 4.3), reflecting the fact that there is poor agreement as to the exact boundaries of ROH: thus one panel may designate one long ROH, whilst another might split it into several shorter ones. This point is also illustrated by the fact that numbers of ROH are more weakly correlated than total length of ROH (table 4.4).

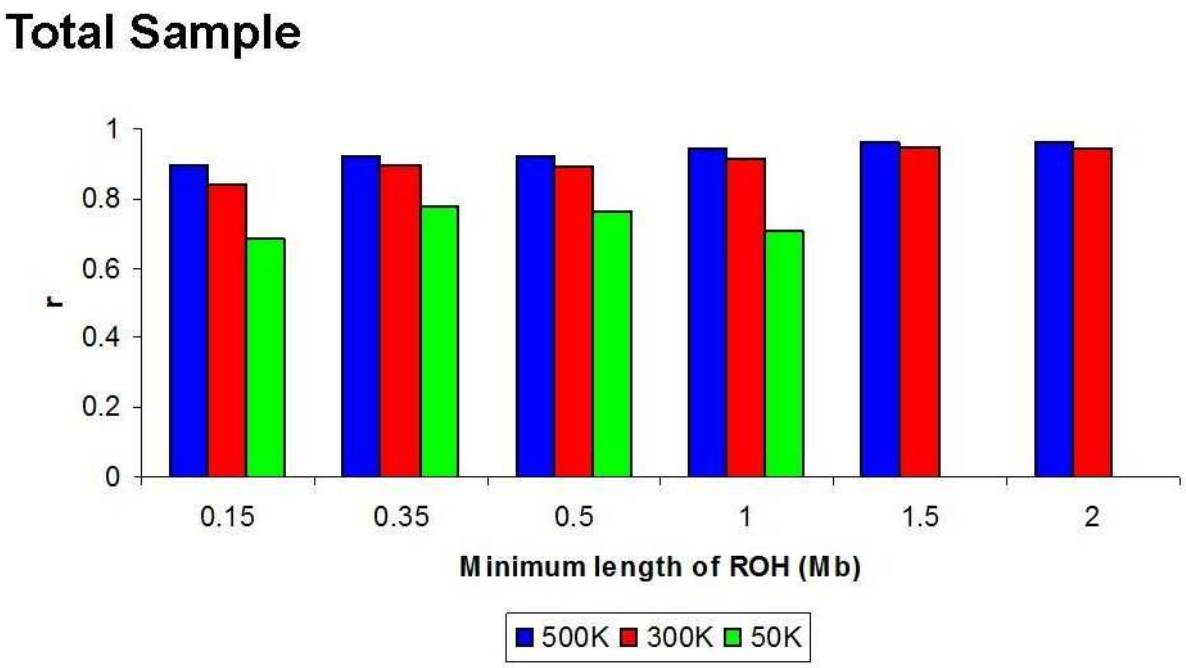
Table 4.3: Correlations in the proportion of the typed autosomal genome in ROH, between the 2400K SNP panel and the 3 other panels, by ROH length category

ROH length (Mb)	Outlier	Panel		
		500K	300K	50K
0.15 - 0.349	**incl**	0.32	0.23	0.15
	excl	0.31	0.20	0.14
0.35 - 0.49	**incl**	0.40	0.23	0.26
	excl	0.40	0.22	0.19
0.5 - 0.99	**incl**	0.63	0.45	0.45
	excl	0.58	0.39	-0.03
1 - 1.49	**incl**	0.70	0.48	0.25
	excl	0.72	0.53	0.09
1.49 - 1.99	**incl**	0.60	0.41	NA
	excl	0.66	0.44	NA

Table 4.4: Correlations in the number of ROH, between the 2400K SNP panel and the other 3 panels, for a range of minimum ROH length thresholds

Minimum ROH length (Mb)	Outlier	Panel		
		500K	300K	50K
0.15	**incl**	0.39	0.32	0.12
0.15	**excl**	0.37	0.29	-0.13
0.35	**incl**	0.61	0.55	0.57
0.35	**excl**	0.57	0.44	0.03
0.5	**incl**	0.71	0.63	0.60
0.5	**excl**	0.63	0.54	-0.03
1	**incl**	0.81	0.69	0.52
1	**excl**	0.80	0.68	0.09
1.5	**incl**	0.83	0.82	NA
1.5	**excl**	0.73	0.63	NA
2	**incl**	0.90	0.87	NA
2	**excl**	0.72	0.57	NA

Figure 4.4: Correlations in the proportion of the typed autosomal genome in ROH, between the 2400K SNP panel and the 3 other panels for a range of minimum ROH length thresholds



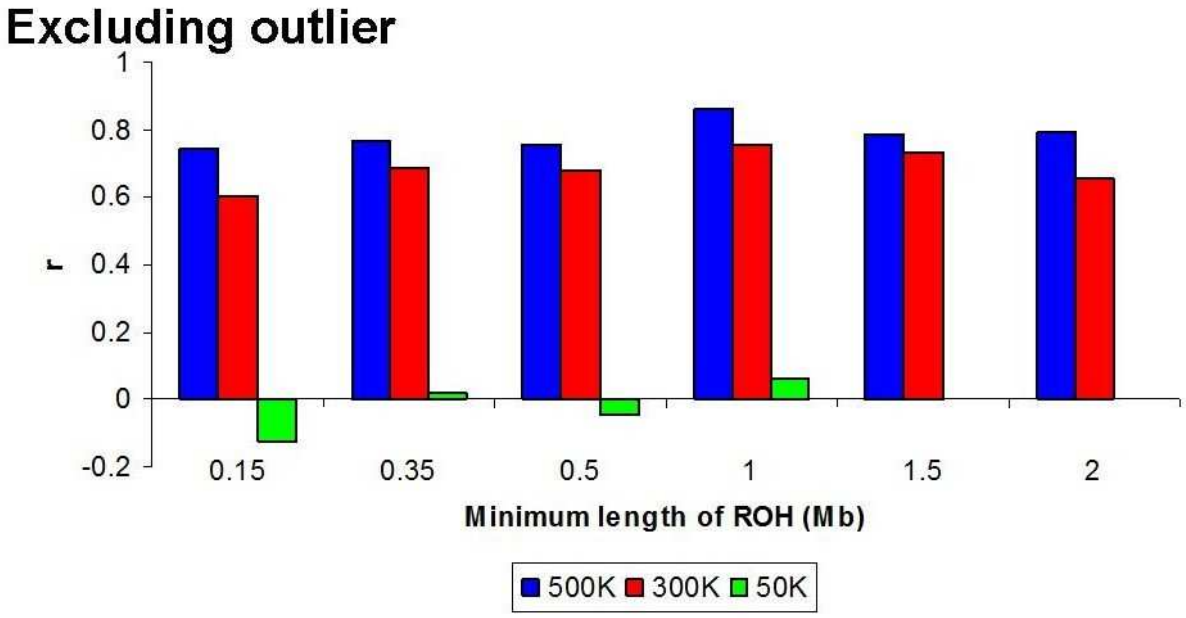
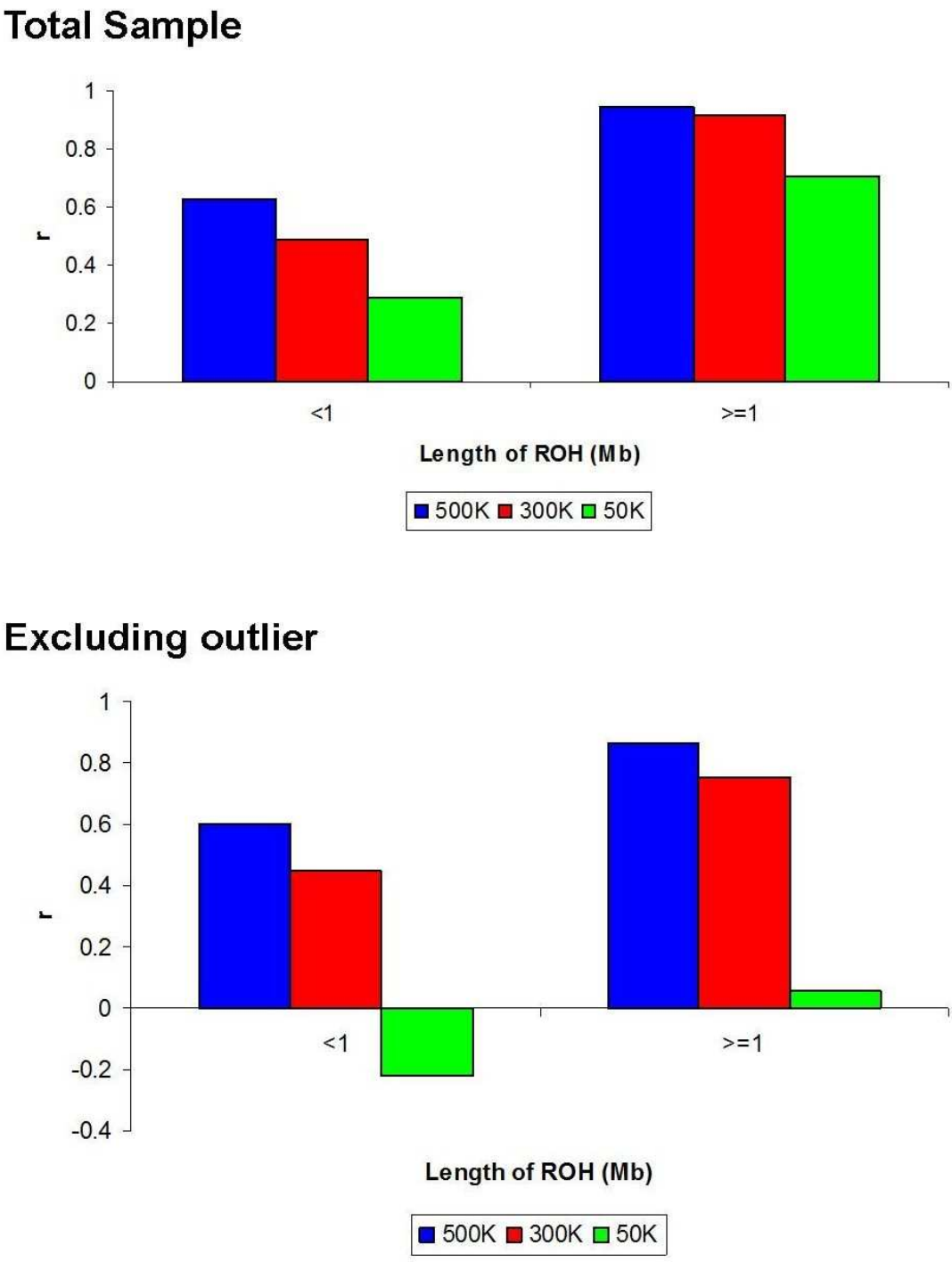


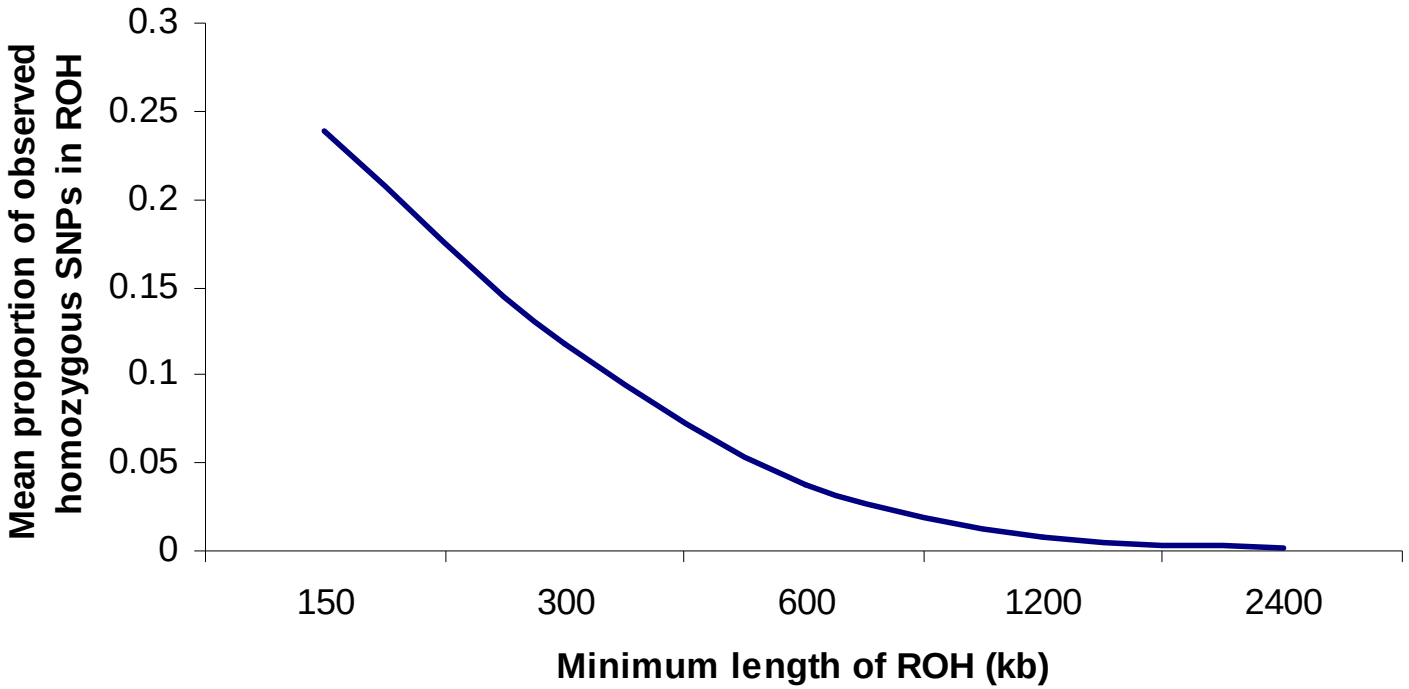
Figure 4.5: Correlations in the proportion of the typed autosomal genome in ROH, between the 2400K SNP panel and the 3 other panels for ROH shorter and longer than 1Mb



4.3.3 Proportion of observed homozygous SNPs in ROH

The proportion of observed homozygous SNPs in ROH of different length categories in the CEU data is shown in Figure 4.6. 24% of observed homozygous SNPs are in ROH longer than 150 kb. This proportion falls to 0.1% in ROH longer than 2400 kb. Only 0.4% are found in ROH longer than 1.5 Mb. Over ¾ of observed homozygous genotypes are found outside ROH longer than 150 kb. This implies that the vast majority of observed homozygous variants are in very short ROH: ROH which are too short to be reliably detected even with much denser SNP panels than those currently available.

Figure 4.6: Mean proportion of observed homozygous SNPs in ROH above a range of minimum length thresholds



4.4 Discussion

The purpose of this analysis was to establish whether or not ROH can be usefully employed to investigate recessivity in non-inbred populations. Three SNP panels of differing densities of SNP coverage were compared with a panel of 2,400,000 SNPs in order to determine the reliability of each panel in detecting ROH of different lengths. It should, of course, be noted that although the 2400K panel is the densest currently available, this has on average only one marker per 1.1 kb and as such is itself an under-estimate of the true picture. Around 20 million SNPs are expected from the forthcoming 1000 Genomes Project.

With this caveat in mind, compared with the 2400K panel, the 500K panel was found to under-estimate the proportion of the typed autosomal genome in ROH longer than 150 kb by over 50%. As the minimum ROH length cut-off is increased, however, panel estimates steadily converge, so that for ROH longer than 1 Mb, there is no significant difference between estimates derived using the 2400K and the 500K panels. For ROH longer than 1.5 Mb, there is no significant difference between the 2400K and the 300K panels. The 50K panel, which was derived by stripping out SNPs in strong LD, is by definition unable to detect ROH arising from short haplotypes inherited as blocks. It is also not dense enough to detect longer ROH and so is not useful for present purposes. SNP panels of this size may still be useful in other organisms with different genetic diversities, patterns of LD or inbreeding, such as domestic animals and their wild or feral relatives.

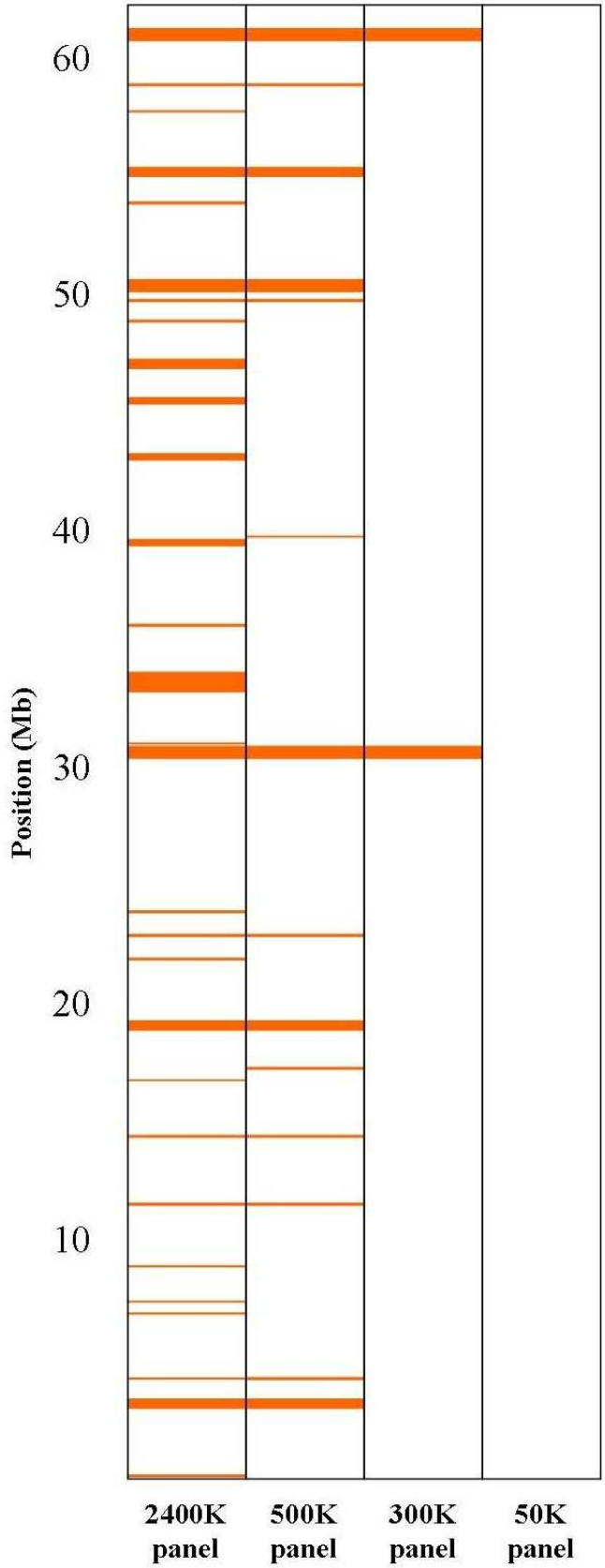
The analysis in chapter 3 showed that F_{ROH} derived using a 300K SNP panel correlated significantly more strongly with F_{ped} than did alternative genomic measures of homozygosity or autozygosity, thus demonstrating that this provides a useful approach for quantifying the effects of inbreeding. This is because the long ROH that are characteristic of recent parental relatedness (longer than 1 – 2 Mb) can be reliably identified using a 300K SNP panel.

The results of the present analysis are consistent with this view and improved correlations between panels when the inbred outlier is included in the sample lend further weight to the view that ROH provide a useful approach for the investigation of inbreeding effects in consanguineous or isolate population samples. Shorter homozygous segments, which are abundant in all individuals, are, however, also a source of individual genetic variation which may be important in disease risk: can existing SNP panels reliably detect these homozygous segments in order to investigate the role of recessivity in disease risk in more cosmopolitan populations?

One approach to answering this question is to investigate how strongly estimates of the percentage of the typed autosomal genome in ROH are correlated when estimated using different panel densities. Whilst both the 500K and 300K panels correlate strongly with the 2400K panel, even when the minimum ROH length threshold is set as low as 150 kb, most of the contribution is coming from the longer ROH. This is illustrated in figure 4.5: correlations between the 2400K and both the 500K and 300K panels for ROH shorter than 1 Mb are only moderate. Neither panel has the density to detect more than a fraction of the short ROH identified by the 2400K

panel. The pictogram in figure 4.7 gives a graphic illustration of the differences in resolution of the 4 panels. This shows chromosome 20 for the outlier (NA12874). There was no particular reason for choosing chromosome 20, except that a short chromosome was preferred because it is easier to see the detail. The longest ROH detected by the 2400K panel for this individual on chromosome 20 measures 610 kb (the apparently longer ROH located at around 33 Mb along the chromosome is actually 2 ROH separated by a narrow gap, which is not easy to see on this diagram). Using the 2400K panel, this individual has 18 ROH measuring 150 – 249 kb on chromosome 20 and 12 ROH measuring over 250 kb. The 500K panel identifies 9 ROH longer than 250 kb but only 5 shorter than 250 kb. The 300K panel identifies only 2 ROH longer than 250 kb (the 2 longest in the sample) but none shorter than 250kb. The 50K panel does not detect any ROH for this individual on chromosome 20. Although it is difficult to see at this resolution, the exact boundaries and numbers of ROH often differ between panels, as these are highly sensitive to panel density. This is demonstrated in tables 4.3 and 4.5: whereas there is strong correlation between panels in the percentage of the typed autosomal genome in ROH longer than a specified minimum ROH length threshold, correlations based on the number of ROH and those based on narrow length thresholds are considerably weaker. This illustration is based on the most inbred individual in the sample: as such, conditions are optimal for the 500K and 300K panels to reveal the ROH. Both panels would fare worse for a chromosome from a non-inbred individual with few long ROH.

Figure 4.7: Diagram of Chromosome 20 for NA12874, showing the length, number and location of ROH according to each panel



This analysis confirms the conclusions of chapter 3: whilst ROH appear to be a sound basis for estimating the effects of (recent) parental inbreeding, currently available SNP panels are not dense enough to be able to enumerate the very short ROH characteristic of outbred subjects. Figure 4.6 suggests that an extremely dense panel would be needed to identify the majority of these ROH. Fewer than 25% of homozygous SNPs in the 2400K panel were found in ROH longer than 150 kb: the vast majority of homozygous SNPs are in very short runs of only a few kilobases to a few tens of kilobases.

Extrapolating from these figures, an estimated 40% of typed homozygous SNPs are predicted to be located in ROH longer than 75 kb and an estimated 60% are predicted to be located in ROH longer than 37.5 kb in populations with similar genomic homozygosity to CEPH. An extremely dense panel would be required to detect the estimated 40% of ROH measuring less than 40 kb. Even the 2400K panel may not be dense enough to enumerate these. Given current technology, quantifying homozygosity in terms of ROH is not, then, likely to be the best approach for investigating recessive effects in outbred samples, although when the 1000 Genomes Project data become available, there may be the potential to update existing SNP panels, at least for use in outbred populations. For the present, however, alternative methods such as H_{pn} are likely to be more fruitful. This can be used in combination with F_{ROH}: whilst H_{pn} is likely to be preferable for investigating general recessive effects, F_{ROH} can be used to stratify the sample into inbred and outbred cohorts and to investigate inbreeding effects. The half Orcadian sample described in chapter 3 can be used to define the maximum percentage of the typed autosomal genome in ROH

for an outbred subject. Combining both approaches in this way allows for separate estimation of the effect on disease or disease traits of both recent inbreeding and homozygosity of more ancient origin. Both approaches will therefore be applied in the following chapters. Chapter 5 is an analysis of recessive effects on a range of biomedically important QT in five genetically isolated populations. Chapter 6 is an analysis of recessive effects in two colorectal cancer case control samples in more cosmopolitan populations.

An immediate future priority beyond the scope of this thesis is to extend the current analysis to include the other three Hapmap founder samples and to repeat it using off-the-shelf Affymetrix SNP chips. A more exploratory aim for the future is to build on work done by Gibson et al (2006) to investigate whether LD maps, which measure the genetic rather than the physical distance between SNPs (Collins, Lau et al. 2004), might bring a useful perspective to ROH data, particularly as the Hap300 and Hap500 panels use SNPs chosen to tag LD blocks throughout the genome. In a recombination cold spot, the physical distance between two SNPs will exceed the genetic distance, whereas in a recombination hotspot the reverse is true. Analysing ROH in terms of genetic, rather than physical, distance is a potentially more sophisticated approach to quantifying the relative age of origin and population prevalence of ROH.

Chapter 5: An investigation of recessive effects in a range of biomedically important quantitative traits in European isolate populations

5.1 Introduction

The purpose of this study is to look for evidence of recessive genetic effects in a range of QTs which play a role in cardiovascular and metabolic disease. Specifically, the percentage of the typed autosomal genome in ROH per individual (F_{ROH} expressed as a percentage for ease of interpretation) is used to assess the effects of recent inbreeding, with an association between trait variation and this measure constituting evidence of recessive effects resulting from recent parental relatedness on the trait in question. Total homozygosity as estimated by H_{pn} is used to assess general recessive effects resulting from all inbreeding loops in an individual's pedigree, however ancient in origin. Study samples are from the five European isolate populations collaborating in the European Special Populations Research Network (EUROSPAN), as described in chapter one. The EUROSPAN populations are ideal for a study of this nature because of increased levels of inbreeding, and therefore homozygosity, compared with outbred populations. It is expected that traits exhibiting dominance variance should be influenced by differences in levels of individual homozygosity.

Heritability analyses in founder populations have shown that systolic blood pressure (SBP), total cholesterol and low-density lipoprotein (LDL) cholesterol have high

dominance variance (Abney, McPeek et al. 2001; Ober, Abney et al. 2001). Weiss et al found results that were in the main consistent with this, except that there was also evidence of dominance variance in high-density lipoprotein (HDL) in males but not in females (Weiss, Pan et al. 2006). Consistent with this, studies in isolated populations with high rates of kin marriage and a large variance in F have found significant associations between trait values for SBP and F measured both by pedigree (Krieger 1969; Martin, Kurczynski et al. 1973; Rudan, Smolej-Narancic et al. 2003; Badaruddoza 2004) and using genomic marker data (Campbell, Carothers et al. 2007). The latter study also found significant associations with DBP, LDL, total cholesterol and Forced Expiratory Flow (FEF) 25.

This study analyses 11 QT underlying cardiovascular disease and metabolic syndrome in the five EUROSPAN populations. In addition, data were available on one of the QT (height) for the colorectal cancer controls enrolled in the SOCCS study described in chapter one.

5.2 Methods

5.2.1 Traits

The following traits were analysed. Units of measurement are shown in brackets.

Lipids

Ln Triglycerides (Ln mmol/L)

Total cholesterol (mmol/L)

HDL cholesterol (mmol/L)

LDL cholesterol (mmol/L)

Hypertension-related

Systolic blood pressure (SBP) (mm Hg)

Diastolic blood pressure (DBP) (mm Hg)

Anthropometry

Height (m)

Weight (kg)

Body Mass Index (BMI) (kg/m^2)

Diabetes-related

Ln Glucose (Ln mmol/L)

Lung function

FVC (L)

Study populations and trait measurement procedures are described elsewhere (Aulchenko, Heutink et al. 2004; Pardo, MacKay et al. 2005; Rudan, Biloglav et al. 2006; Campbell, Carothers et al. 2007; Pattaro, Marroni et al. 2007; Tenesa, Farrington et al. 2008; Johansson, Marroni et al. 2009). The average (SD) age of each population was 56 (15) for CROAS (range 18 – 93); 53 (15) for ERF (range 18 – 92); 45 (16) for MICROS (range 18 – 87); 47 (21) for NSPHS (range 14 – 91); 54 (16) for ORCADES (range 17 – 97) and 52 (6) for SOCCS (range 21 – 61). Because 18.3% of the total sample was taking treatment for hypertension and 8.3% for high cholesterol, systolic and diastolic blood pressure values were adjusted for those on anti-hypertensive medication and total cholesterol, HDL, LDL and triglyceride values were adjusted for those being treated with lipid-lowering drugs. The literature on the effects of anti-hypertensive medication on blood pressure and on the effects of

lipid lowering therapies on cholesterol was examined in order to make appropriate trait adjustments. This is summarised in tables 5.1 and 5.2.

Table 5.1: Literature on the effects of anti-hypertensive medication on blood pressure

Study	Sample details	Findings
(Tobin, Sheehan et al. 2005)	Simulations and 659 men aged 60-74 (120 of whom taking antihypertensives)	Found reduction of 10 mm Hg in SBP for 1 treatment. Suggests using 10-15 mm Hg reduction to take account of the additive effects of taking >1 treatment
(Law, Wald et al. 2003)	Meta-analysis – 40,000 treated, 16,000 placebo	All categories of antihypertensives produced similar reductions in BP. Mean = 9.1 mm Hg (SBP) and 5.5 mm Hg (DBP). Combining different categories of drug had an additive effect on BP reduction.
(Turnbull, Neal et al. 2008)	Meta-analysis >190,000 people	SBP reduction of up to 7.2 mm Hg for under 65s and up to 9.3 mm Hg for over 65s. DBP reduction of 2-4 mm Hg. NB not directly comparable because this study compared each medication either with placebo or with a less intensive BP medication.
(Cui, Hopper et al. 2003)	2912 people in 767 families (244 on treatment). Most only on 1 treatment	Reduction of 10 mm Hg in SBP and 5 mm Hg in DBP

The Tobin, Law and Cui papers are consistent with each other estimating a 10 mm Hg reduction in SBP with one antihypertensive treatment. The estimated reduction in DBP values is around half the reduction in SBP. Law suggests an additive effect for each additional category of antihypertensive. The Turnbull study is the biggest and most recent. Estimates appear to be lower than for the other three studies; however this study is not directly comparable to the others and is difficult to interpret because each drug was compared either with a placebo or with a less intensive blood pressure treatment.

Because 51% of those taking anti-hypertensive medication in the present study were taking more than one category of anti-hypertensive drug, 15 mm Hg were added to

raw SBP values and 7.5 mm Hg to raw DBP values of all those on medication, regardless of how many different medications being taken. Another approach would be to add 10 mm Hg to those on 1 medication, 20 to those on 2 and so on, but because precise drugs, dosages and compliance are unknown, there is a danger that this would be adding arbitrary precision, particularly when inter-individual response to these drugs varies so much.

Table 5.2: Literature on the effects of lipid-lowering therapy on cholesterol

Author	Sample details	Findings
(LaRosa, He et al. 1999)	Meta-analysis >30,000 people	20% reduction in total cholesterol 28% reduction in LDL cholesterol 5% increase in HDL 13% reduction in triglycerides Mean duration of treatment = 5.4 years
(Baigent, Keech et al. 2005)	Meta-analysis > 90,000 people	28% reduction in LDL at 1 year, 21% at 5 years (reduction probably because of non-compliance)

Two large meta-analyses reporting on the effects on lipids of lipid lowering therapies were examined (table 5.2). As a result, total cholesterol values were increased by 20%, LDL by 28%, triglycerides by 13% and HDL values were decreased by 5% for those on medication.

Subjects taking diabetes medication were removed from the glucose analysis, as there were only 7% on treatment and so removing them resulted in very little loss in power. Distributions of trait values were examined to see if they conformed to approximate normality. Glucose and triglycerides were severely skewed and so were log transformed.

5.2.2 Genotyping

All five EUROSPAN samples were genotyped using the Illumina Infinium HumanHap300 platform (Illumina, San Diego), as described for the ORCADES sample in chapter 3. Full descriptions of genotyping for the other EUROSPAN samples and for SOCCS are given elsewhere (Johansson, Vavruch-Nilsson et al. 2005; Liu, Arias-Vasquez et al. 2007; Pattaro, Marroni et al. 2007; Tenesa, Farrington et al. 2008; Vitart, Rudan et al. 2008). Taking each population individually, SNPs with more than 10% missing were removed, as were SNPs failing HWE at $p < 0.0001$ and SNPs with MAF < 0.01. Full QC procedures to remove individuals with low genotyping and discrepant pedigree and genomic data had already been applied, as described in the above references. A consensus SNP panel of 288,598 autosomal SNPs was then created, consisting of SNPs satisfying the above QC criteria in all six populations. Final sample numbers were 722 SOCCS (height only), 718 ORCADES, 789 CROAS, 1097 MICROS, 642 NSPHS and 881 ERF.

5.2.3 Measures of homozygosity

Three different measures of homozygosity were employed: two F_{ROH} measures to assess the effects of recent inbreeding, plus H_{pn} to assess general homozygosity effects attributable to more distant parental relatedness. All were estimated using PLINK (Purcell, Neale et al. 2007), using the parameters described in chapter 3, except that instead of expressing F_{ROH} as a *proportion* of the typed autosomal genome in ROH, these measures are expressed as a *percentage* of the typed autosomal genome in ROH. This is to simplify interpretation of the results, as

numbers are very small and thus more difficult to grasp when expressed as a proportion. Two different F_{ROH} measures are used: $F_{ROH1.5}$ is the percentage of the typed autosomal genome in ROH greater than or equal to 1.5 Mb and F_{ROH5} is the equivalent measure for ROH greater or equal to 5 Mb. As described in chapters 3 and 4, there are two reasons for using a cut-off of 1.5 Mb. Firstly, as shown in chapter 3, all individuals in all populations surveyed have ROH shorter than 1.5 Mb; differences between individuals with known parental inbreeding in their pedigrees start to become apparent for ROH longer than 1.5 Mb. Secondly, as shown in chapter 4, a 300,000 SNP panel is highly reliable for the identification of ROH longer than 1.5 Mb but increasingly unreliable at identifying ROH shorter than this. The 5 Mb threshold is used in addition to the 1.5 Mb threshold because 100% of individuals in the NSPHS and over 98% of individuals overall in the other population samples had ROH longer than 1.5 Mb: thus a higher threshold may differentiate better between individuals in such inbred samples.

5.2.4 Statistical analysis

Because all five EUROSPAN studies are family-based, with high levels of relatedness between individuals, subjects are not independent and therefore conventional regression techniques are not valid. For this reason, EUROSPAN data were analysed in GenABEL (available at http://mga.bionet.nsc.ru/nlru/GenABEL/ (Aulchenko, Ripke et al. 2007)) using a linear mixed polygenic model based on the trait, specified covariates and a genomic kinship matrix. This kinship matrix estimates pairwise relatedness, derived on the basis of IBS sharing, weighted by allele frequency, so that a pair of individuals sharing a rare allele is estimated to be

more closely related than a pair sharing a common allele. The consensus SNP panel described above was used to generate the kinship matrix. Age and sex were fitted as fixed effects. The model also estimates narrow sense heritability (h^2). The SOCCS sample consists of unrelated individuals, so data were analysed in SPSS using simple linear regression, with the trait as the dependent variable and age, sex and each homozygosity measure in turn fitted as independent variables.

For each trait, each population was analysed in turn. Results were then combined in a meta-analysis using the inverse variance method to combine effect size estimates from each sample (Aulchenko 2008). This weights each sample estimate by the inverse of the squared standard error of the regression coefficient, so that the smaller the standard error of the study, the greater the contribution it makes to the pooled regression coefficient.

Let β_i be the regression coefficient, and s_i^2 be the squared standard error of the regression coefficient, of N studies, where:

$i \in 1,2,...,N$

Let w_i be the weight of an individual sample, defined as:

$w_i = 1/\ s_i^2$

The pooled regression coefficient is defined as:

$$\beta = \frac{\sum_{i=1}^{N} w_i \beta_i}{\sum_{i=1}^{N} w_i}$$

5.2.5 Socio-economic status

Socio-economic status is a potential confounder of the association between homozygosity and height, and indeed many other QT, since both reduced height and inbreeding are known to be associated with low socio-economic status (Mackenbach 1992). Data on socio-economic status were only available for the SOCCS and ORCADES data sets. These samples were split by deprivation category using the Carstairs scores for Scottish post-code sectors derived by Carstairs and Morris (Carstairs and Morris 1990; McLoone 2001). This area-based measure takes z-scores of four variables (male unemployment, households with no car, overcrowding and head of household's social class) and converts them into 1 -7 categories, with category 1 the least deprived. Mean F_{ROH} and H_{pn} values by deprivation category were compared to assess the risk of confounding by socio-economic status.

5.3 Results

5.3.1 F_{ROH} and H_{pn} in EUROSPAN samples

Table 5.3 shows sample means and variances for $F_{ROH1.5}$, F_{ROH5} and H_{pn}. In all cases, variances are an order of magnitude higher in the isolate population samples than in the outbred SOCCS sample. Variances for the NSPHS sample are much higher than the other isolate samples. Table 5.4 shows mean QT values, split by sex, for each population. Figure 5.1 shows the proportion of each sample with ROH longer than 0.5, 1.5 and 5 Mb. All individuals in all samples have ROH longer than 0.5 Mb. All individuals in the NSPHS sample have ROH longer than 1.5 Mb and almost ¾ have ROH longer than 5 Mb. Over 98% of individuals in the other four samples have

ROH longer than 1.5 Mb and approaching ½ have ROH longer than 5 Mb (52 % of ERF sample, 47% of MICROS and CROAS samples and 41% of ORCA sample).

Table 5.3: Mean (variance) of F_{ROH} and H_{pn} by sample

Sample	N	$F_{ROH1.5}$	F_{ROH5}	H_{pn}
CROAS	789	0.90 (0.93)	0.39 (0.44)	0.654 (2.11×10^{-5})
ERF	881	1.04 (1.15)	0.48 (0.60)	0.653 (2.35×10^{-5})
MICROS	1097	0.86 (0.91)	0.38 (0.44)	0.652 (2.05×10^{-5})
NSPHS	642	2.76 (5.60)	1.32 (2.78)	0.661 (10.9×10^{-5})
ORCADES	718	0.76 (0.68)	0.29 (0.32)	0.652 (1.79×10^{-5})
SOCCS	722	0.27 (0.05)	0.02 (0.02)	0.651 (0.42×10^{-5})

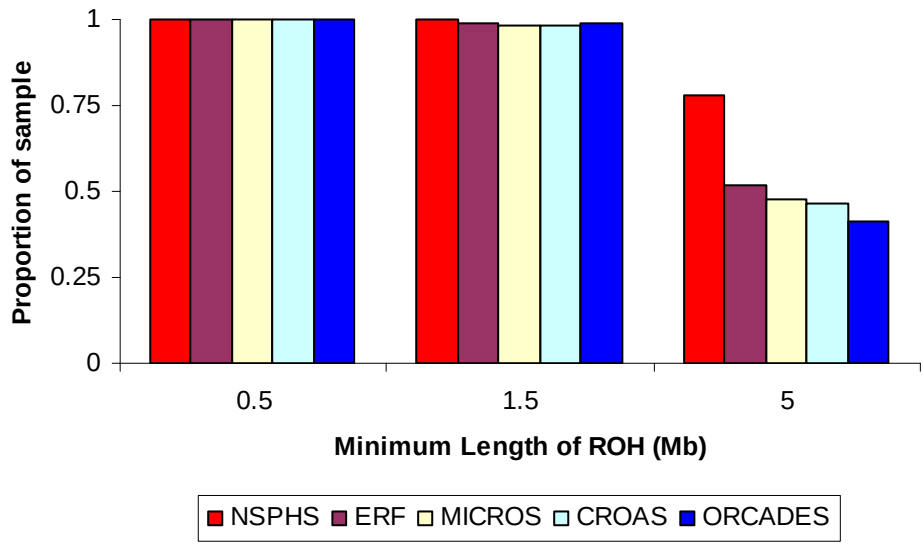
Table 5.4: Mean Trait values by sex and sample

Trait	Sample	Females			Males		
		N	Mean	SE	N	Mean	SE
Total cholesterol	CROAS	453	5.19	0.05	323	4.99	0.06
	ERF	510	5.75	0.05	318	5.63	0.06
	MICROS	615	5.99	0.05	471	5.90	0.06
	NSPHS	340	5.97	0.07	299	5.96	0.08
	ORCADES	385	6.05	0.06	332	5.77	0.06
HDL	CROAS	452	1.12	0.01	322	1.09	0.01
	ERF	510	1.34	0.02	318	1.11	0.02
	MICROS	615	1.81	0.01	471	1.53	0.01
	NSPHS	340	1.72	0.02	299	1.46	0.02
	ORCADES	385	1.81	0.02	332	1.50	0.02
LDL	CROAS	452	3.36	0.05	322	3.08	0.05
	ERF	509	3.84	0.05	314	3.82	0.05
	MICROS	615	3.53	0.05	471	3.68	0.05
	NSPHS	340	3.55	0.06	299	3.67	0.06
	ORCADES	385	3.64	0.06	332	3.72	0.06
Ln triglycerides	CROAS	454	0.36	0.02	323	0.49	0.03
	ERF	509	0.15	0.02	318	0.33	0.03
	MICROS	615	0.14	0.02	471	0.34	0.03
	NSPHS	340	0.51	0.03	299	0.76	0.04
	ORCADES	385	0.17	0.02	332	0.24	0.03
SBP	CROAS	456	141	1.4	322	142	1.4
	ERF	515	142	1.1	319	150	1.2
	MICROS	621	132	0.9	474	137	0.9
	NSPHS	338	124	1.3	293	128	1.2
	ORCADES	373	130	1.2	324	137	1.1
DBP	CROAS	456	81	0.6	322	84	0.7
	ERF	515	81	0.5	319	85	0.6
	MICROS	621	80	0.5	474	81	0.5
	NSPHS	338	75	0.5	293	77	0.5
	ORCADES	372	77	0.6	323	79	0.6
Height	CROAS	456	1.62	0.003	322	1.76	0.004
	ERF	490	1.60	0.003	299	1.73	0.004
	MICROS	612	1.61	0.003	467	1.73	0.003
	NSPHS	339	1.58	0.004	299	1.71	0.004
	ORCADES	373	1.61	0.003	324	1.75	0.004
	SOCCS	352	1.62	0.004	370	1.76	0.004
Weight	CROAS	442	71.0	0.6	322	85.7	0.7
	ERF	490	68.1	0.6	299	82.5	0.8
	MICROS	612	65.2	0.5	468	78.4	0.6
	NSPHS	337	65.0	0.7	298	78.2	0.8
	ORCADES	373	71.0	0.7	324	85.7	0.7
BMI	CROAS	442	27.2	0.2	322	27.6	0.2
	ERF	490	26.5	0.2	299	27.7	0.3
	MICROS	612	25.3	0.2	467	26.1	0.2
	NSPHS	335	26.0	0.3	298	26.7	0.3
	ORCADES	373	27.4	0.3	324	28.1	0.2

Table 5.4 continued: Mean Trait values by sex and sample

Trait	Sample	Females			Males		
		N	Mean	SE	N	Mean	SE
FVC	CROAS	445	3.7	0.05	321	5.3	0.07
	ORCADES	371	3.1	0.03	323	4.2	0.05
Ln glucose	CROAS	404	1.68	0.0078	300	1.71	0.0088
	ERF	478	1.49	0.0072	295	1.56	0.0095
	MICROS	600	1.51	0.0050	452	1.55	0.0059
	ORCADES	378	1.64	0.0051	320	1.70	0.0070

Figure 5.1: Proportion of sample with ROH longer than specified length, by population



Mean $F_{ROH1.5}$ is shown in figure 5.2 and mean F_{ROH5} is shown in figure 5.3 for each sample, with 95% confidence intervals. For both F_{ROH} measures, NSPHS means are significantly higher than mean values for any of the other samples, which suggests that the prevalence of recent parental relatedness is higher in the NSPHS sample than in other samples. The mean F_{ROH5} value for ORCADES is significantly lower than all other samples and the mean $F_{ROH1.5}$ value for ORCADES is significantly lower

than all except MICROS, which suggests that the prevalence of recent parental relatedness is lower in the ORCA sample than in the other samples.

Sample means with 95% confidence intervals for H_{pn} are shown in figure 5.4. Again, the NSPHS sample mean is significantly higher than the means of the other samples. CROAS is significantly higher than MICROS, ERF and ORCA. ERF and ORCA have overlapping confidence intervals and MICROS is significantly lower than all other samples.

Figure 5.2: Mean percentage of typed autosomal genome in ROH longer than 1.5 Mb ($F_{ROH1.5}$), with 95% confidence intervals, by population

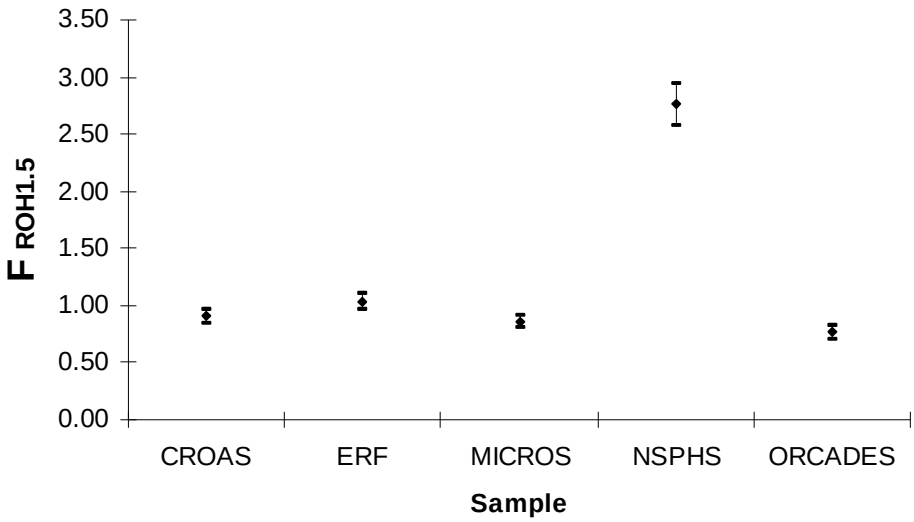


Figure 5.3: Mean percentage of typed autosomal genome in ROH longer than 5 Mb (F_{ROH5}), with 95% confidence intervals, by population

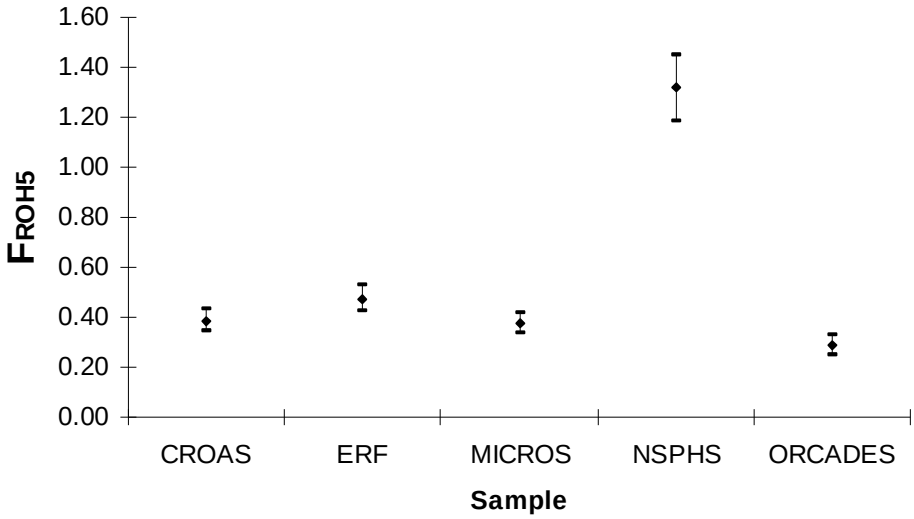
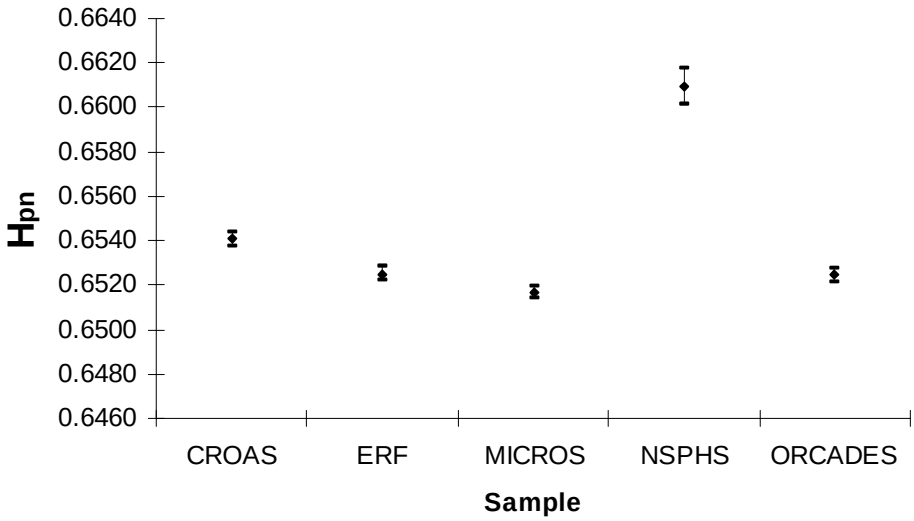


Figure 5.4: Mean H_{pn}, with 95% confidence intervals, by population



5.3.2 Analysis of QT

Narrow sense heritabilities (h^2) adjusted by age and sex and estimated using GenABEL are shown in table 5.5. Results of the regression analyses are shown in tables 5.6 to 5.10. Heritability estimates are imprecise in all samples, but particularly in CROAS and ORCADES. This reflects the number of relative pairs available for analysis. Estimates are most precise for the MICROS sample. Significant differences between populations were found for LDL and total cholesterol. The MICROS estimate, which is the most precise, is significantly higher than all the other estimates except for ORCADES (and 95% confidence intervals for MICROS and ORCADES only just overlap). The ORCADES estimate of h^2 for LDL is significantly higher than the NSPHS estimate and 95% confidence intervals only just overlap between ORCADES and ERF. Results are similar for total cholesterol. There are no significant differences between samples for the other QT examined, although 95% confidence intervals for height only just overlap between NSPHS (which has the most precise and the lowest heritability estimate for height) and ERF.

Table 5.5: Heritability estimates with 95% confidence interval by trait and population, controlled for age and sex

Trait	CROAS	ERF	ORCADES	NSPHS	MICROS
ln glucose	0.27 (0.11 to 0.43)	0.48 (0.30 to 0.66)	0.18 (0.02 to 0.34)	N/a	0.32 (0.23 to 0.41)
ln triglycerides	0.38 (0.18 to 0.58)	0.29 (0.15 to 0.43)	0.36 (0.19 to 0.53)	0.37 (0.25 to 0.49)	0.38 (0.28 to 0.48)
total cholesterol	0.29 (0.09 to 0.49)	0.30 (0.17 to 0.43)	0.49 (0.33 to 0.65)	0.23 (0.13 to 0.33)	0.71 (0.60 to 0.82)
HDL	0.32 (0.11 to 0.53)	0.50 (0.35 to 0.65)	0.48 (0.32 to 0.64)	0.59 (0.44 to 0.74)	0.43 (0.32 to 0.54)
LDL	0.22 (0.01 to 0.43)	0.24 (0.12 to 0.36)	0.53 (0.36 to 0.70)	0.23 (0.13 to 0.33)	0.77 (0.67 to 0.87)
SBP	0.40 (0.14 to 0.66)	0.30 (0.17 to 0.43)	0.29 (0.14 to 0.44)	0.16 (0.01 to 0.31)	0.21 (0.13 to 0.29)
DBP	0.25 (0.04 to 0.46)	0.33 (0.19 to 0.47)	0.39 (0.24 to 0.54)	0.31 (0.17 to 0.45)	0.29 (0.22 to 0.36)
FVC	0.35 (0.12 to 0.58)	N/a	0.37 (0.20 to 0.54)	N/a	N/a
height	0.90 (0.66 to 1.00)	0.89 (0.76 to 1.00)	0.91 (0.60 to 1.00)	0.68 (0.60 to 0.76)	0.83 (0.71 to 0.95)
weight	0.50 (0.26 to 0.74)	0.48 (0.34 to 0.62)	0.47 (0.28 to 0.66)	0.53 (0.41 to 0.65)	0.65 (0.53 to 0.77)
BMI	0.39 (0.10 to 0.68)	0.47 (0.33 to 0.61)	0.50 (0.31 to 0.69)	0.45 (0.31 to 0.59)	0.55 (0.42 to 0.68)

Table 5.6: Results of regression analysis (lipid traits)

Trait	Sample	Measure	β	SE	p-value	95% CI
Total cholesterol	**CROAS**	$F_{ROH1.5}$	-0.018	0.046	ns	-0.11 to 0.07
		F_{ROH5}	-0.039	0.15	ns	-0.33 to 0.25
		H_{pn}	1.8	4.7	ns	-7 to 11
	ERF	$F_{ROH1.5}$	0.095	0.053	ns	-0.01 to 0.20
		F_{ROH5}	0.11	0.096	ns	-0.08 to 0.30
		H_{pn}	20	13	ns	-6 to 46
	MICROS	$F_{ROH1.5}$	-0.015	0.047	ns	-0.11 to 0.08
		F_{ROH5}	-0.037	0.052	ns	-0.14 to 0.07
		H_{pn}	-4.6	11	ns	-26 to 17
	NSPHS	$F_{ROH1.5}$	0.0082	0.013	ns	-0.018 to 0.034
		F_{ROH5}	-0.0088	0.017	ns	-0.043 to 0.025
		H_{pn}	3.2	4.1	ns	-5 to 11
	ORCADES	$F_{ROH1.5}$	0.027	0.038	ns	-0.05 to 0.10
		F_{ROH5}	0.049	0.065	ns	-0.08 to 0.18
		H_{pn}	1.9	11	ns	-20 to 24
HDL	**CROAS**	$F_{ROH1.5}$	0.0044	0.066	ns	-0.13 to 0.13
		F_{ROH5}	0.0083	0.012	ns	-0.016 to 0.033
		H_{pn}	0.2	0.76	ns	-1.3 to 1.7
	ERF	$F_{ROH1.5}$	0.015	0.011	ns	-0.007 to 0.037
		F_{ROH5}	0.02	0.017	ns	-0.012 to 0.052
		H_{pn}	2.4	3.2	ns	-3.9 to 8.7
	MICROS	$F_{ROH1.5}$	0.011	0.0099	ns	-0.008 to 0.030
		F_{ROH5}	0.0061	0.011	ns	-0.015 to 0.028
		H_{pn}	1.34	3.2	ns	-4.9 to 7.5
	NSPHS	$F_{ROH1.5}$	0.0031	0.006	ns	-0.009 to 0.015
		F_{ROH5}	-0.00068	0.003	ns	-0.0066 to 0.0053
		H_{pn}	1.2	2.8	ns	-4.2 to 6.6
	ORCADES	$F_{ROH1.5}$	0.017	0.018	ns	-0.019 to 0.053
		F_{ROH5}	0.026	0.026	ns	-0.025 to 0.077
		H_{pn}	3.8	3.8	ns	-4 to 11
LDL	**CROAS**	$F_{ROH1.5}$	-0.0064	0.026	ns	-0.058 to 0.045
		F_{ROH5}	-0.031	0.086	ns	-0.20 to 0.14
		H_{pn}	-1.1	7.8	ns	-16 to 14
	ERF	$F_{ROH1.5}$	0.071	0.058	ns	-0.04 to 0.19
		F_{ROH5}	0.082	0.11	ns	-0.13 to 0.30
		H_{pn}	16	13	ns	-9 to 41
	MICROS	$F_{ROH1.5}$	-0.07	0.054	ns	-0.18 to 0.36
		F_{ROH5}	-0.095	0.32	ns	-0.72 to 0.53
		H_{pn}	-15	13	ns	-40 to 10
	NSPHS	$F_{ROH1.5}$	-0.012	0.049	ns	-0.11 to 0.08
		F_{ROH5}	-0.026	0.033	ns	-091 to 0.039
		H_{pn}	-1.4	4.9	ns	-11 to 8
	ORCADES	$F_{ROH1.5}$	1.3	2.9	ns	-4.3 to 6.9
		F_{ROH5}	0.026	0.049	ns	-0.07 to 0.12
		H_{pn}	-2.9	22	ns	-45 to 39

Table 5.6 continued: Results of regression analysis (lipid traits)

Trait	Sample	Measure	β	SE	p-value	95% CI
ln triglycerides	**CROAS**	$F_{ROH1.5}$	-0.015	0.11	ns	-0.22 to 0.19
		F_{ROH5}	-0.016	0.071	ns	-0.16 to 0.12
		H_{pn}	3.8	2.8	ns	-1.8 to 9.3
	ERF	$F_{ROH1.5}$	0.017	0.017	ns	-0.017 to 0.051
		F_{ROH5}	0.016	0.029	ns	-0.041 to 0.073
		H_{pn}	3.4	4	ns	-4 to 11
	MICROS	$F_{ROH1.5}$	0.007	0.029	ns	-0.049 to 0.063
		F_{ROH5}	0.0053	0.012	ns	-0.019 to 0.030
		H_{pn}	2.4	35	ns	-67 to 71
	NSPHS	$F_{ROH1.5}$	0.023	0.016	ns	-0.009 to 0.055
		F_{ROH5}	0.029	0.024	ns	-0.017 to 0.075
		H_{pn}	4.6	4.3	ns	-4 to 13
	ORCADES	$F_{ROH1.5}$	0.0075	0.022	ns	-0.035 to 0.050
		F_{ROH5}	0.012	0.031	ns	-0.049 to 0.073
		H_{pn}	-0.27	1.6	ns	-3.5 to 2.9

No significant association was found in any population between homozygosity and any of the lipid traits, regardless of the measure used.

Table 5.7: Results of regression analysis (hypertension related traits)

Trait	Sample	Measure	β	SE	p-value	95% CI
SBP	**CROAS**	$F_{ROH1.5}$	1.9	1	ns	-0.2 to 3.9
		F_{ROH5}	2.8	1.5	ns	-0.2 to 5.9
		H_{pn}	193	239	ns	-276 to 662
	ERF	$F_{ROH1.5}$	-1.1	0.98	ns	-3.0 to 0.8
		F_{ROH5}	-1.6	1.4	ns	-4.4 to 1.2
		H_{pn}	-280	246	ns	-763 to 203
	MICROS	$F_{ROH1.5}$	-1.01	2.2	ns	-5.2 to 3.2
		F_{ROH5}	-1.6	1.4	ns	-4.2 to 1.1
		H_{pn}	-175	324	ns	-812 to 462
	NSPHS	$F_{ROH1.5}$	1.60E-03	0.016	ns	-0.030 to 0.034
		F_{ROH5}	0.093	0.21	ns	-0.31 to 0.50
		H_{pn}	-15	34	ns	-81 to 51
	ORCADES	$F_{ROH1.5}$	-1.7	0.9	ns	-3.49 to 0.05
		F_{ROH5}	-2.6	1.3	0.04	-5.1 to -0.1
		H_{pn}	-219	199	ns	-609 to 171
DBP	**CROAS**	$F_{ROH1.5}$	0.072	0.19	ns	-0.31 to 0.45
		F_{ROH5}	-0.015	0.15	ns	-0.31 to 0.28
		H_{pn}	-34	88	ns	-206 to 138
	ERF	$F_{ROH1.5}$	0.035	0.12	ns	-0.21 to 0.28
		F_{ROH5}	0.018	0.13	ns	-0.23 to 0.27
		H_{pn}	-26	48	ns	-121 to 69
	MICROS	$F_{ROH1.5}$	-0.73	0.48	ns	-1.7 to 0.2
		F_{ROH5}	-0.97	1.1	ns	-3.2 to 1.2
		H_{pn}	-147	118	ns	-378 to 84
	NSPHS	$F_{ROH1.5}$	0.051	0.14	ns	-0.23 to 0.33
		F_{ROH5}	0.12	0.43	ns	-0.73 to 0.97
		H_{pn}	19	24	ns	-28 to 66
	ORCADES	$F_{ROH1.5}$	-0.39	0.35	ns	-1.1 to 0.3
		F_{ROH5}	-0.85	0.56	ns	-1.9 to 0.2
		H_{pn}	35	88	ns	-137 to 207

One significant association was found with a blood pressure trait. A reduction of 2.6 mm Hg was associated with a 1% increase in F_{ROH5} in the ORCADES sample ($p = 0.04$).

Table 5.8: Results of regression analysis (anthropometric traits)

Trait	Sample	Measure	β	SE	p-value	95% CI
Height	**CROAS**	$F_{ROH1.5}$	-0.0019	0.0019	ns	-0.0057 to 0.0019
		F_{ROH5}	-0.006	0.008	ns	-0.022 to 0.010
		H_{pn}	-0.86	0.8	ns	-2.4 to 0.7
	ERF	$F_{ROH1.5}$	-0.0083	0.014	ns	-0.037 to 0.020
		F_{ROH5}	-0.0098	0.0098	ns	-0.029 to 0.009
		H_{pn}	-1.7	1.1	ns	-3.8 to 0.4
	MICROS	$F_{ROH1.5}$	-0.0046	0.0021	0.026	-0.0087 to -0.0006
		F_{ROH5}	-0.0055	0.0026	0.038	-0.0107 to -0.0003
		H_{pn}	-0.62	0.3	0.041	-1.21 to -0.03
	NSPHS	$F_{ROH1.5}$	-0.0022	0.0012	ns	-0.0045 to 0.0001
		F_{ROH5}	-0.0041	0.001	<0.0001	-0.0062 to -0.0021
		H_{pn}	-1.4	0.27	<0.0001	-1.9 to -0.9
	ORCADES	$F_{ROH1.5}$	-0.0022	0.002	ns	-0.0061 to 0.0017
		F_{ROH5}	-0.0039	0.0037	ns	-0.011 to 0.003
		H_{pn}	-0.63	0.38	ns	-1.4 to 0.1
Weight	**CROAS**	$F_{ROH1.5}$	0.57	0.4	ns	-0.2 to 1.4
		F_{ROH5}	0.61	0.61	ns	-0.6 to 1.8
		H_{pn}	54	59	ns	-61 to 169
	ERF	$F_{ROH1.5}$	-0.17	0.23	ns	-0.63 to 0.29
		F_{ROH5}	0.098	0.2	ns	-0.30 to 0.50
		H_{pn}	-62	76	ns	-212 to 88
	MICROS	$F_{ROH1.5}$	-1.1	1.3	ns	-3.7 to 1.5
		F_{ROH5}	-1.5	1.3	ns	-4.0 to 1.0
		H_{pn}	-200	173	ns	-539 to 139
	NSPHS	$F_{ROH1.5}$	-0.76	0.47	ns	-1.7 to 0.2
		F_{ROH5}	-0.75	0.47	ns	-1.7 to 0.2
		H_{pn}	-190	234	ns	-648 to 268
	ORCADES	$F_{ROH1.5}$	-0.0094	0.094	ns	-0.19 to 0.17
		F_{ROH5}	0.031	0.22	ns	-0.39 to 0.46
		H_{pn}	20	58	ns	-93 to 133
BMI	**CROAS**	$F_{ROH1.5}$	0.34	0.16	0.033	0.03 to 0.65
		F_{ROH5}	0.41	0.25	ns	-0.07 to 0.89
		H_{pn}	45	31	ns	-15 to 105
	ERF	$F_{ROH1.5}$	0.22	0.14	ns	-0.05 to 0.49
		F_{ROH5}	0.36	0.2	ns	-0.02 to 0.74
		H_{pn}	35	27	ns	-19 to 89
	MICROS	$F_{ROH1.5}$	-0.27	0.29	ns	-0.84 to 0.30
		F_{ROH5}	-0.4	0.24	ns	-0.88 to 0.08
		H_{pn}	-58	61	ns	-179 to 63
	NSPHS	$F_{ROH1.5}$	-0.095	0.094	ns	-0.28 to 0.09
		F_{ROH5}	-0.14	0.13	ns	-0.39 to 0.11
		H_{pn}	-22	22	ns	-66 to 22
	ORCADES	$F_{ROH1.5}$	0.046	0.14	ns	-0.22 to 0.32
		F_{ROH5}	0.12	0.56	ns	-1.0 to 1.2
		H_{pn}	25	91	ns	-152 to 202

Table 5.9: Results of regression analysis in SOCCS data set (height only)

Measure	β	SE	p-value	95% CI
$F_{ROH1.5}$	-0.002	0.011	ns	-0.024 to 0.020
F_{ROH5}	0.014	0.02	ns	-0.025 to 0.053
H_{pn}	-1.66	1.23	ns	-4.1 to 0.8

A reduction in height was significantly associated with increased homozygosity using all three measures in the MICROS sample (p-values ranging from 0.03 to 0.04) and for F_{ROH5} and H_{pn} in the NSPHS sample ($p < 0.0001$). A reduction of ~ 0.5 cm in height was associated with an increase of 1% in F_{ROH}. The direction of the effect in the ORCADES, CROAS and ERF samples were consistent with this; however results were non-significant. The direction of effect in the SOCCS sample was also consistent with this for H_{pn} and for $F_{ROH1.5}$, but not for F_{ROH5}. The latter is unreliable in the SOCCS sample as, being predominantly outbred, only a handful of subjects have ROH longer than 5 Mb. One further significant association was found for anthropometric traits: an increase of 0.34 kg/m2 was associated with a 1% increase in $F_{ROH1.5}$ in the CROAS sample ($p = 0.03$).

Table 5.10: Results of regression analysis (other traits)

Trait	Sample	Measure	β	SE	p-value	95% CI
FVC	**CROAS**	$F_{ROH1.5}$	-0.029	0.035	ns	-0.097 to 0.039
		F_{ROH5}	-0.042	0.051	ns	-0.14 to 0.06
		H_{pn}	-1.2	12	ns	-25 to 22
	ORCADES	$F_{ROH1.5}$	-0.063	0.031	0.045	-0.125 to -0.001
		F_{ROH5}	-0.092	0.047	ns	-0.1848 to 0.0007
		H_{pn}	-11	6.6	ns	-24 to 2
Ln glucose	**CROAS**	$F_{ROH1.5}$	0.01	0.024	ns	-0.036 to 0.056
		F_{ROH5}	0.0068	0.0087	ns	-0.010 to 0.024
		H_{pn}	2.8	2.4	ns	-2.0 to 7.6
	ERF	$F_{ROH1.5}$	0.0023	0.0073	ns	-0.012 to 0.017
		F_{ROH5}	0.002	0.02	ns	-0.037 to 0.041
		H_{pn}	0.83	1.3	ns	-1.6 to 3.3
	MICROS	$F_{ROH1.5}$	0.0055	0.0044	ns	-0.003 to 0.014
		F_{ROH5}	0.0059	0.006	ns	-0.006 to 0.018
		H_{pn}	1.4	0.91	ns	-0.4 to 3.1
	ORCADES	$F_{ROH1.5}$	-0.0055	0.006	ns	-0.018 to 0.006
		F_{ROH5}	-0.0054	0.0092	ns	-0.023 to 0.013
		H_{pn}	-0.58	1.1	ns	-2.7 to 1.5

One significant result was found for FVC: in the ORCADES sample, a reduction of 0.06 L was associated with a 1% increase in $F_{ROH1.5}$ ($p = 0.045$).

5.3.3 Meta-analysis

Data were then meta-analysed as described in section 5.2.4 above. Two alternative Bonferroni adjustments to correct for multiple testing were considered. If data are adjusted on the basis of 11 different traits, the adjusted p-value is 0.0045 (for a nominal threshold of 0.05). If a stricter adjustment is made on the basis of 52 different tests, the adjusted p-value is 0.00096.

The only trait to remain significantly associated with homozygosity after meta-analysis and correction for multiple testing is height. With the less strict Bonferroni correction, height remains significantly associated with all three homozygosity

measures. With the stricter correction for multiple testing, H_{pn} and F_{ROH5} remain highly significantly associated with height.

Table 5.11: Meta-analysis

Trait	Measure	β_{pooled}	SE_{pooled}	p_{pooled}	95% CI
BMI	$F_{ROH1.5}$	0.050	0.061	ns	-0.07 to 0.17
	F_{ROH5}	0.0086	0.090	ns	-0.17 to 0.18
	H_{pn}	8.0	14	ns	-20 to 36
DBP	$F_{ROH1.5}$	0.0032	0.080	ns	-0.15 to 0.16
	F_{ROH5}	-0.021	0.093	ns	-0.20 to 0.16
	H_{pn}	4.6	20	ns	-35 to 44
FVC	$F_{ROH1.5}$	-0.048	0.023	0.041	-0.094 to -0.002
	F_{ROH5}	-0.069	0.035	0.047	-0.1368 to -0.0009
	H_{pn}	-8.7	5.8	ns	-20 to 3
HDL	$F_{ROH1.5}$	0.0075	0.0045	ns	-0.001 to 0.016
	F_{ROH5}	0.0011	0.0028	ns	-0.0044 to 0.0066
	H_{pn}	0.53	0.68	ns	-0.8 to 1.9
Height	$F_{ROH1.5}$	-0.0025	0.00082	0.0019	-0.0042 to -0.0009
	F_{ROH5}	-0.0043	0.00093	0.000004	-0.0061 to -0.0025
	H_{pn}	-0.99	0.17	0.000000005	-1.3 to -0.7
LDL	$F_{ROH1.5}$	-0.0068	0.020	ns	-0.046 to 0.032
	F_{ROH5}	-0.0075	0.025	ns	-0.057 to 0.042
	H_{pn}	-1.0	3.7	ns	-8.3 to 6.3
Ln glucose	$F_{ROH1.5}$	0.0019	0.0031	ns	-0.0042 to 0.0081
	F_{ROH5}	0.0035	0.0042	ns	-0.005 to 0.012
	H_{pn}	0.73	0.59	ns	-0.4 to 1.9
Ln triglycerides	$F_{ROH1.5}$	0.016	0.0098	ns	-0.003 to 0.035
	F_{ROH5}	0.011	0.0097	ns	-0.008 to 0.030
	H_{pn}	1.4	1.3	ns	-1.1 to 3.9
SBP	$F_{ROH1.5}$	0.0012	0.016	ns	-0.031 to 0.033
	F_{ROH5}	0.0038	0.20	ns	-0.39 to 0.40
	H_{pn}	-23	32	ns	-86 to 41
Total cholesterol	$F_{ROH1.5}$	0.011	0.011	ns	-0.011 to 0.033
	F_{ROH5}	-0.0052	0.016	ns	-0.036 to 0.025
	H_{pn}	2.9	2.8	ns	-2.6 to 8.3
Weight	$F_{ROH1.5}$	-0.033	0.084	ns	-0.20 to 0.13
	F_{ROH5}	0.0069	0.137895	ns	-0.26 to 0.28
	H_{pn}	0.91	35	ns	-68 to 70

Most of the signal for the association between homozygosity and height comes from the NSPHS sample (although the MICROS sample also shows a significant association and the direction of effect is consistent across all studies). The h^2 estimate for height was, however, lower in NSPHS than in the other samples. For these reasons, the meta-analyses for height were done both with and without the NSPHS sample. Without the NSPHS sample, results remain significant, but not when adjusted for multiple testing (table 5.12).

Table 5:12 Meta-analysis for height, excluding NSPHS

Measure	β_{pooled}	SE_{pooled}	p_{pooled}	95% CI
$F_{ROH1.5}$	-0.00286	0.0011	0.012	-0.0051 to -0.0006
F_{ROH5}	0.00503	0.002	0.013	-0.0090 to -0.0010
H_{pn}	-0.72	0.22	0.001	-1.15 to -0.29

5.3.4 Homozygosity effects adjusted for recent inbreeding

Both F_{ROH} measures used here estimate the effects of recent parental inbreeding (F_{ROH5} estimating more recent inbreeding than $F_{ROH1.5}$). H_{pn} estimates the total effects of inbreeding in an individual's entire ancestral demographic history, from the very ancient shared parental ancestry common to all of us, to the very recent. In order to assess the effect on height of homozygosity of ancient origin (or in other words, H_{pn} controlled for inbreeding), the analysis was repeated with age, sex and each F_{ROH} measure in turn fitted as fixed effects. The only sample where there was evidence that homozygosity controlled for inbreeding was significantly associated with a reduction in height was the NSPHS sample (table 5.13). When the results were meta-analysed, the p-value obtained was higher than for the NSPHS sample

alone, despite the increased sample size, and when the meta-analysis was repeated excluding the NSPHS sample, the significant association disappeared (table 5.14).

Table 5.13: H_{pn} and height, controlled for age, sex and recent inbreeding, by sample

Fixed Effect	Sample	β	SE	p-value	95% CI
$F_{ROH1.5}$	**NSPHS**	-2.3	0.48	<0.0001	-3.2 to -1.4
	ORCADES	-0.78	0.61	ns	-2.0 to 0.4
	ERF	-0.54	0.50	ns	-1.5 to 0.4
	CROAS	-0.4	0.44	ns	-1.3 to 0.5
	MICROS	0.83	0.70	ns	-0.5 to 2.2
	SOCCS	-1.9	1.4	ns	-4.6 to 0.8
F_{ROH5}	**NSPHS**	-3.5	0.63	<0.0001	-4.7 to -2.3
	ORCADES	-0.57	0.46	ns	-1.5 to 0.3
	ERF	-1.3	1.0	ns	-3.3 to 0.7
	CROAS	-0.5	0.55	ns	-1.6 to 0.6
	MICROS	0.11	0.23	ns	-0.3 to 0.6
	SOCCS	0.014	0.02	ns	-0.03 to 0.05

Table 5:14: H_{pn} and height, controlled for age, sex and recent inbreeding – meta-analysis

Sample	Fixed effect	β_{pooled}	SE_{pooled}	p_{pooled}	95% CI
Total sample	$F_{ROH1.5}$	-0.83	0.23	0.0003	-1.3 to -0.4
	F_{ROH5}	0.0089	0.20	ns	-0.4 to 0.4
Excluding NSPHS	$F_{ROH1.5}$	-0.39	0.26	ns	-0.90 to 0.12
	F_{ROH5}	0.012	0.02	ns	-0.027 to 0.051

5.3.5 Socio-economic status

Data on socio-economic status were available for the ORCADES and SOCCS samples only. On the basis of Carstairs deprivation scores, there is negligible variation in deprivation in the ORCADES sample: 99.6% of the sample is classified as deprivation category 3. There is more variation in the SOCCS sample, so mean values for H_{pn} and $F_{ROH1.5}$ were calculated for each deprivation category. No significant difference by deprivation category was found using either homozygosity measure (figures 5.5 and 5.6).

Figure 5.5: Mean (95% confidence interval) $F_{ROH1.5}$ by Carstairs deprivation category (SOCCS sample)

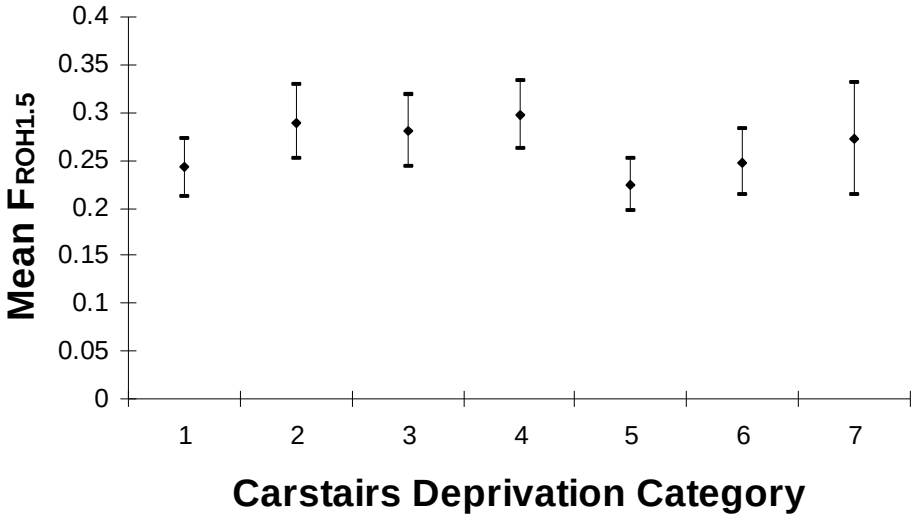
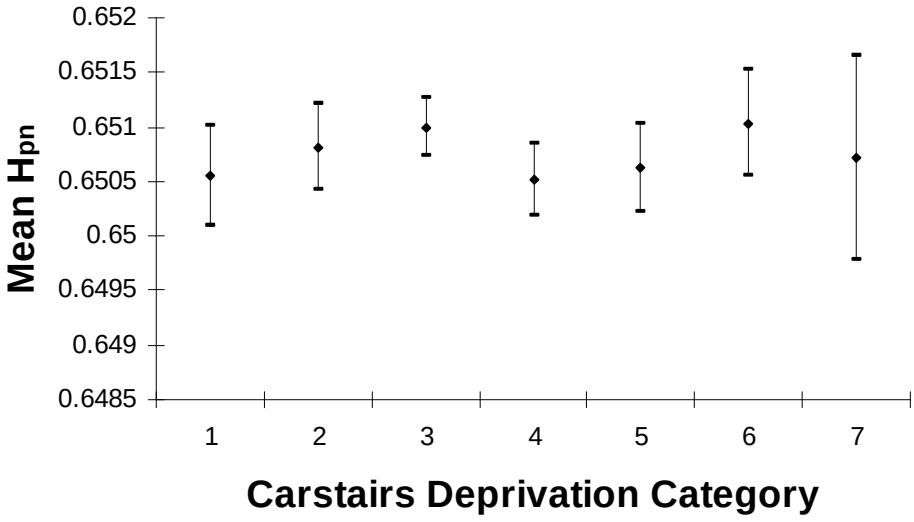


Figure 5.6: Mean (95% confidence interval) H_{pn} by Carstairs deprivation category (SOCCS sample)



5.4 Discussion

5.4.1 Population characteristics

This study was based on 5 isolate populations characterised by higher mean levels of parental relatedness than are generally found in European populations. One more cosmopolitan European sample was included in the analysis of height. Four of the five isolate samples are very similar to each other in their distribution of ROH (figures 5.1 to 5.4 and table 5.3); however the fifth (NSPHS from Northern Sweden) is very different. On average, NSPHS subjects have almost 3 times the proportion of their genome in ROH compared with subjects from the other samples (figures 5.2 and 5.3) and significantly higher H_{pn} values (figure 5.4). There are a number of reasons for this. Most obviously, levels of consanguineous marriage may be higher in NSPHS than in the other study populations. Secondly, NSPHS has lower N_e, as a result of very long term isolation: the other isolate populations are by comparison relatively recently isolated. Finally, in terms of allele frequencies, this population, which consists of individuals with Saami and non-Saami heritage, is genetically more distant than the others are from the western European populations used in the discovery of the majority of SNPs making up the Hap300 array. Ascertainment bias may therefore be an issue because allele frequencies will in general be more different and as such may influence the probability of homozygosity. The NSPHS sample is also interesting because it consists of two distinct sub-groups, one with Saami and one with non-Saami heritage. Analysis by sub-population would reveal the extent to which this signal is coming from one or both groups.

In the other isolate samples, differences between F_{ROH} and H_{pn} statistics are interesting. CROAS is not significantly different to ERF or MICROS for mean F_{ROH5} (i.e. levels of recent parental relatedness are no higher in CROAS than in ERF or MICROS); however mean H_{pn} is significantly higher for CROAS than for all the other samples except NSPHS, suggesting a smaller effective population size in this island population deep in the past. Results for ORCADES are similar: F_{ROH} statistics suggest that levels of recent parental relatedness are significantly lower than in all the other samples; however mean H_{pn} is significantly higher than for MICROS, suggesting smaller long term N_e than for the MICROS sample.

5.4.2 Heritability estimates

The only traits where significant differences were found between populations were LDL and total cholesterol, where the MICROS estimate was significantly higher than all except the ORCADES estimate. This may in part reflect true population differences in heritability: environmental influences on trait values differ widely across the study populations, ranging from a Mediterranean diet and climate in the CROAS sample to a sub-Arctic climate and a lifestyle characterised by very high physical activity and a diet high in game and fish in NSPHS. Differences in heritability estimates may also be to some extent an artefact of the methodology used. Estimates are very imprecise, particularly in the CROAS and ORCADES samples, reflecting lower numbers of relative pairs available to estimate heritability. Results may also be skewed if traits are sexually dimorphic. One approach to this problem would be to analyse males and females separately; however using GenABEL, this would mean that heritabilities would be estimated by comparing

males only with males and females only with females. This can lead to very variable results, as the number of pairs used in the analysis is cut quite dramatically, especially in small sample sizes such as these. Thus a few pairs could be very influential, leading to unreliable results (*personal communication*, Veronique Vitart). For this reason, it was decided on balance to control for sex rather than analyse males and females separately.

5.4.3 Lipid and Blood Pressure traits

A high dominance variance has been reported in systolic blood pressure and LDL cholesterol in the Hutterites (Abney, McPeek et al. 2001). For this reason, there is a theoretical expectation that these QT will be influenced by inbreeding. Consistent with this, various studies have found evidence of significant positive association between blood pressure and $F_{ped.}$ Thirteen studies investigating the association between blood pressure and inbreeding or consanguinity were found. Three were discarded because they were ecological studies which did not use an individual measure of inbreeding. Of the remaining 10 studies, 9 measured inbreeding using F_{ped}, of which 5 used categorical measures (e.g. inbred v not inbred) (Soyannwo, Kurashi et al. 1998; Saleh, Mahfouz et al. 2000; Badaruddoza 2004; Bener, Hussain et al. 2006; Rudan, Biloglav et al. 2006) and 4 used F_{ped} as a continuous measure (Krieger 1969; Martin, Kurczynski et al. 1973; Eldon, Axelsson et al. 2001; Rudan, Smolej-Narancic et al. 2003). Only 1 study (Campbell, Carothers et al. 2007) used a genomic measure of inbreeding. These studies are summarised in table 5.15. Four found no evidence for an association between blood pressure and inbreeding (Martin, Kurczynski et al. 1973; Soyannwo, Kurashi et al. 1998; Eldon, Axelsson et al. 2001;

Bener, Hussain et al. 2006), although 2 of these were relatively small and so may have been under powered (Martin, Kurczynski et al. 1973; Eldon, Axelsson et al. 2001). Five found evidence of an association between inbreeding and blood pressure (Krieger 1969; Martin, Kurczynski et al. 1973; Saleh, Mahfouz et al. 2000; Rudan, Smolej-Narancic et al. 2003; Badaruddoza 2004; Rudan, Biloglav et al. 2006; Campbell, Carothers et al. 2007).

Six studies examining the association between inbreeding and cholesterol were found. One was an ecological study, which did not use an individual measure of inbreeding, and so was excluded. The remaining 5 are summarised in table 5.16. Four used pedigree measures of inbreeding (Martin, Kurczynski et al. 1973; Eldon, Axelsson et al. 2001; Rudan, Biloglav et al. 2006; Isaacs, Sayed-Tabatabaei et al. 2007) and 1 used a genomic measure, as described above (Campbell, Carothers et al. 2007). One study found no significant association between inbreeding and cholesterol (Rudan, Biloglav et al. 2006). One study found a significant negative correlation between HDL and inbreeding (Eldon, Axelsson et al. 2001). One study found a significant negative association between cholesterol (presumably total cholesterol, although the paper does not make this clear) and inbreeding in males under 20 but a significant positive association in males over 40 (Martin, Kurczynski et al. 1973). Two papers found significant positive associations between both total cholesterol and LDL cholesterol and inbreeding (Campbell, Carothers et al. 2007; Isaacs, Sayed-Tabatabaei et al. 2007).

Table 5.15 Summary of literature on blood pressure and inbreeding

Reference	Measure of inbreeding/homozygosity	Sample	Findings	Effect size
Soyannwo et al, 1998	Categorical (offspring of 1st cousins, 2nd cousins, more distant cousins, unrelated parents)	5671 residents of Buraidah, Gassim Region, Saudi Arabia (includes both urban and rural areas)	No evidence of association between consanguinity category and blood pressure	NA
Rudan, Biloglav et al, 2006	Ordered categorical: • inbred = evidence of inbreeding from pedigree or isonymy; • autochthonous = all 4 grandparents from subject's village of residence; • admixed = mother's parents born in one village and father's parents born in another; • outbred = at least 3 grandparents born in different larger Croatian mainland settlements	Location: Dalmatian islands, Croatia. 76 subjects in each category (total 304) Study conducted in 2002.	Significant association between ordered categories and SBP (p=0.006). This remained when inbred and autochthonous categories were combined into an inbred category and admixed and outbred categories were combined into an outbred category (p=0.005)	Mean (interquartile range) SBP for inbred category = 150 (40), and for outbred category = 140 (35)
Saleh, 2000	Categorical (consanguineous vs. non-consanguineous, but not clear how consanguinity defined)	1312 primary school children aged 6-10, Kuwait.	Prevalence of hypertension much higher in children of consanguineous parents	OR (95% CI) of hypertension = 1.61 (1.11 – 3.56)
Bener, 2006	Categorical (any degree of consanguinity found from pedigrees (mostly parents related as 2nd cousins or closer) vs. non-consanguineous).	876 Qatari females age 15+. Study conducted 2004-2005	No significant difference in OR of hypertension between offspring of consanguineous vs. non-consanguineous subjects	NA

Table 5.15 continued: Summary of literature on blood pressure and inbreeding

Reference	Measure of inbreeding/homozygosity	Sample	Findings	Effect size
Badaruddoza, 2004	Categorical (consanguineous vs. non-consanguineous, based on 3 ancestral generations)	3253 Muslim North Indian children aged 6-14 (one child per family), Aligarh District, Uttar Pradesh	DBP and SBP significantly higher in both males and females in consanguineous category ($p<0.001$)	Mean SBP was 4mm Hg higher in consanguineous than in non-consanguineous group. Mean DBP was 5 mm Hg higher in males in consanguineous group compared with non-consanguineous group and 7mm Hg higher in females in consanguineous group compared with non-consanguineous
Eldon, 2001	Continuous (F_{ped} based on unspecified number of generations)	119 Icelandic people living in Iceland and Canada.	No significant association between F_{ped} and blood pressure	NA
Rudan, Smolej-Narancic et al 2003	Continuous (F_{ped} based on 4-5 ancestral generations)	2760 adults resident in Croatian island isolate settlements, Dalmatia. Data collected 1979-81.	Significant association between F_{ped} and both SBP and DBP	SBP was 20 mmHg higher in offspring of 1st cousins compared with offspring of unrelated parents.
Martin, 1973	Continuous (F_{ped} based on complete pedigrees back to 1800 and incomplete back to 1700)	489 subjects from S-Leut Hutterites, a large religious isolate in USA and Canada.	No association with DBP, some association with SBP in some sub-groups	NA
Krieger, 1969	Continuous (F_{ped} based on unspecified number of generations)	3465 children from Sao Paolo, Brazil	Significant inbreeding effect ($p<0.01$) on DBP	10% increase in F_{ped} associated with increase in DBP of 35mm Hg
Campbell, 2007	Genomic measure of inbreeding (relative heterozygosity = excess/expected heterozygosity derived from panel of 1240 microsatellite markers and cross checked with pedigree data)	385 adults resident in Croatian island isolates	SBP and DBP both significantly associated with relative heterozygosity ($p<0.05$)	SBP was 6.8 mmHg higher and DBP was 3.3 mmHg higher in offspring of 1st cousins compared with offspring of unrelated parents.

Table 5.16: Summary of literature on cholesterol and inbreeding

Reference	Measure of inbreeding/homozygosity	Sample	Findings
Rudan, Biloglav et al, 2006	Ordered categorical: • inbred = evidence of inbreeding from pedigree or isonymy; • autochthonous = all 4 grandparents from subject's village of residence; • admixed = mother's parents born in one village and father's parents born in another; • outbred = at least 3 grandparents born in different larger Croatian mainland settlements	Location: Dalmatian islands, Croatia. 76 subjects in each category (total 304) Study conducted in 2002.	No evidence of significant association, although borderline significant associations for total cholesterol and HDL.
Eldon, 2001	Continuous (F_{ped} based on unspecified number of generations)	119 Icelandic people living in Iceland and Canada.	HDL and F_{ped} significantly negatively correlated ($p<0.05$)
Martin, 1973	Continuous (F_{ped} based on complete pedigrees back to 1800 and incomplete back to 1700)	489 subjects from S-Leut Hutterites, a large religious isolate in USA and Canada.	Inbreeding significantly associated with reduction in cholesterol in males aged <20 and an increase in cholesterol in males aged 40+
Isaacs, 2007	Categorical (F_{ped} based on >15 generation pedigree data divided into quartiles)	868 subjects from Dutch isolate	Inbreeding significantly associated with total cholesterol ($p_{trend} = 0.02$) and LDL cholesterol ($p_{trend} = 0.05$)
Campbell, 2007	Genomic measure of inbreeding (relative heterozygosity = excess/expected heterozygosity derived from panel of 1240 microsatellite markers and cross checked with pedigree data)	385 adults resident in Croatian island isolates	Total cholesterol was 6.8% and LDL was 9.6% higher in offspring of 1st cousins compared with offspring of unrelated parents.

The present study found no evidence for an inbreeding effect on either blood pressure or lipid traits. Part of the explanation for this, particularly for blood pressure, might be the high level of treatment for hypertension in the study samples. In order to account for the effects of treatment, blood pressure values of those being treated for hypertension were adjusted as described above. An examination of meta-analyses investigating the effects on SBP and DBP of various hypertensive regimes found that each hypertensive medication reduced SBP by around 10 mm Hg and DBP by around 5 mm Hg, with the effect of multiple medications being additive (Law, Wald et al. 2003). Following this logic, and taking into account the numbers of participants taking 1, 2, 3 and 4 different medications, SBP was adjusted upwards by 15 mm Hg and DBP upwards by 7.5 mm Hg for those being treated for hypertension. This approach is, however, far from ideal: the meta-analyses were all based on clinical trial data but there is a very great difference between clinical trial conditions and real life; there is a great deal of individual variation in treatment response which cannot be predicted; finally, 1 of the meta-analyses seemed to suggest a more modest response to treatment, although results were difficult to interpret.

Adjusting lipid trait values for those being treated with statins is similarly problematic. Variation in individual response, dosage, compliance and duration of treatment and the differences between clinical trial conditions and real life make any adjustment far from ideal. An obvious response to this problem would be to remove all those being treated for hypertension or raised cholesterol from the analysis. Unfortunately, given the large numbers of subjects being treated for hypertension,

this would dramatically reduce the sample size. It would also have the effect of reducing trait variance. The prevalence of anti-hypertensive treatment was almost certainly lower in the studies carried out longer ago or in the developing world, and was effectively zero in the Dalmatian study of Rudan et al (2003).

It is important to note that there are many potential inaccuracies in the measurement of phenotypic data. In addition to measurement error within individual populations, differences in measurement protocols among the study populations may have introduced bias. Data on self-reported medication may be affected by recall bias and coding of drugs may be prone to error. All of these types of error may impact on the results of the analysis.

Only one other study used a genomic measure to estimate inbreeding (Campbell, Carothers et al. 2007). The other studies cited here all used F_{ped}, based on pedigree data of varying quality and completeness. It is difficult to say how this might have affected results, except to say that the expectation is that if an inbreeding effect is detectable with a very imperfect pedigree-based measure, the expectation is that it should be easier to detect with an enhanced genomic measure.

An important difference between this study and all other studies examined is that this was the only one to correct for relatedness within the sample. GenABEL uses a genomic kinship matrix to correct for closer trait resemblance between relative pairs than between pairs of unrelated individuals. Failure to correct in this way may have resulted in false positive results in other studies. To demonstrate this, LDL was

analysed using simple linear regression, adjusting for age and sex but not kinship. Whilst estimated effect sizes were the same as given in table 5.6, p-values were lower. Using this approach, H_{pn} was significantly associated with LDL in ERF (p = 0.02) and significantly negatively associated with LDL in MICROS (p = 0.035).

Another explanation for the failure of this study to find an association with lipid or blood pressure traits might be that many of the cited studies were conducted in highly consanguineous populations, where levels of inbreeding were likely to be much higher than in the EUROSPAN samples. For example, the study by Saleh et al reports the rate of consanguineous marriage in Kuwait, from which the study sample is drawn, to be 54.3% and the mean population F_{ped} to be 0.02 (Saleh, Mahfouz et al. 2000). Even some of the Dalmatian studies were conducted in villages with reported mean population F_{ped} of this order (up to 0.05 for one village) (Rudan, Smolej-Narancic et al. 2003). Publication bias, favouring studies reporting evidence for association, may also be a factor.

An examination of confidence intervals, both by individual sample (tables 5.6 and 5.7) and in the meta-analysed total sample (table 5.11), suggests that insufficient power in the present study is unlikely to be the reason why associations were not found here between homozygosity and blood pressure or lipid traits. If the direction of effect was consistent across samples and if confidence intervals were very wide or were heavily skewed in either a positive or negative direction, this would suggest that increased sample sizes might produce statistically significant results. This, however, is not the case: confidence intervals in the meta-analysed sample are

narrow and because there is inconsistency among populations (for example, beta values for SBP are negative in ERF and ORCADES but positive in CROAS; beta values for LDL are positive in ERF but negative in MICROS) confidence intervals are not skewed away from zero in the meta-analysed sample. Given that we are only able to detect a small proportion of the homozygosity present, it remains an open question whether ROH influence these traits in the general population.

5.4.4 Height

The present study found evidence for significant inbreeding depression associated with height. Most of this result was driven by a very high association between reduced height and both H_{pn} and F_{ROH} in the NSPHS sample, although results were also significant in the MICROS sample and the estimated effect size and direction of effect was consistent across all samples. The high p-value in the NSPHS sample may be attributable to increased power to detect an effect because of higher levels of parental relatedness compared with the other samples. The higher p-value in the MICROS sample compared with CROAS, ERF and ORCADES may reflect increased power to detect an effect because of the larger sample size. Nevertheless, the fact that results in the NSPHS sample were so much higher than those found in the other EUROSPAN samples means that it may be prudent to consider these results both including and excluding NSPHS. Excluding NSPHS, results remained significant at the 0.05 level, however not when adjusted for multiple testing. The literature on height and inbreeding was searched in order to put these findings in context. Eight observational studies investigating inbreeding and height were found. Two found no evidence of inbreeding depression on height (Neel, Schull et al. 1970;

Campbell, Carothers et al. 2007) whilst six found evidence of a reduction in height associated with inbreeding (Morton 1958; Martin, Kurczynski et al. 1973; 1983; Krishan 1986; Badaruddoza 2004; Zottarelli, Sunil et al. 2007). These studies are summarised in table 5.16. All of the studies that found clear evidence of inbreeding depression involved children or babies (Morton 1958; Freire-Maia 1983; Krishan 1986; Badaruddoza 2004; Zottarelli, Sunil et al. 2007): it is possible that these results reflect differences in growth rate rather than height, since adult height is the result not only of childhood growth, but also of loss of height during ageing (Weedon, Lango et al. 2008). The two studies involving adults (Martin, Kurczynski et al. 1973; Campbell, Carothers et al. 2007) found either no or ambiguous evidence of inbreeding depression. Another point that should be taken into account when evaluating these is socio-economic status, a known confounder of height, which may have biased the results of two of the studies examined (Morton 1958; Badaruddoza 2004).

Table 5.17: Summary of literature on inbreeding and height

Reference	Measure of inbreeding/homozygosity	Sample	Findings	Effect size
Campbell, 2007	Genomic measure of inbreeding (relative heterozygosity = excess/expected heterozygosity derived from panel of 1240 microsatellite markers and cross checked with pedigree data)	385 adults resident in Croatian island isolates.	No evidence of inbreeding effect. Results controlled for socio-economic status (SES).	NA
Martin, 1973	Continuous (F_{ped} based on complete pedigrees back to 1800 and incomplete back to 1700)	489 subjects from S-Leut Hutterites, a large religious isolate in USA and Canada.	Significant evidence of inbreeding depression in some age-sex groups but significant increase in height associated with inbreeding in others. SES not measured but this is a communal population.	Insufficient data provided by authors
Krishan, 1986	F_{ped} categorical (consanguineous v non-consanguineous derived from 3-5 generation pedigrees)	502 Sheikh Sunni Muslim boys age 11-16, Delhi.	Significant reduction in height at all ages in offspring of consanguineous parents. No difference between groups in father's income, occupation and education.	0.3 – 2.5% difference in height between consanguineous and non-consanguineous groups (data presented by year of age)
Badaruddoza, 2004	F_{ped} categorical (F_{ped} =0 and F_{ped} >0)	1443 North Indian children, age 6-14. SES not evaluated.	Significant reduction in height in F_{ped} >0 group ($p<0.001$). SES not measured, so may bias results.	Mean height reduction of 3.75 cm (males) and 5.27 cm (females)
Neel et al, 1969	Categorised by type of parental cousin relationship	1343 middle school children, Japan. Data collected 1965.	No evidence of inbreeding depression. Results controlled for SES.	NA

Table 5.17 continued: Summary of literature on inbreeding and height

Reference	Measure of inbreeding/homozygosity	Sample	Findings	Effect size
Zottarelli et al, 2007	Categorised into no relation, 1st cousin, 2nd cousin, other relation	10,194 children under 5 years old, Egypt. Data collected 2000.	"Stunting" (based on z score indicators of height for age) significantly higher in offspring of consanguineous parents. Parents' education and other social factors evaluated to avoid confounding.	Offspring of 1st cousins significantly ($p<0.01$) higher odds of being below -2SD for height for age.
Morton et al, 1958	F_{ped} categorical (more remote than 2nd cousin parental relationship counted as unrelated)	~75,000 babies age 8-9 months, Japan. Data collected 1948-54.	Very small but significant reduction in height associated with inbreeding. SES not measured, so may bias results.	Mean height in unrelated group was 2.3mm taller than mean height in the offspring of 1st cousins group
Freire-Maia, 1983	F_{ped}	534 south Brazilian school children. Data collected 1964-65.	Reduction in height significantly associated with inbreeding. Father's occupation evaluated to avoid confounding.	2cm decrease in height associated with 10% increase in F_{ped}.

There is, then, some empirical evidence of an inbreeding effect on growth and/or adult height. How consistent is this with what is known about the genetics of height? Evidence of inbreeding depression on a trait is consistent with high dominance variance; however there is little published evidence of dominance variance for height. A recent study by Weedon et al (2008) found no strong evidence of deviation from an additive genetic model for height. Visscher et al (2007) did not find significant evidence of dominance variance, although the authors state that dominance variance was difficult to assess in their sib-pair design because genomewide additive and dominance coefficients were highly correlated. Three heritability studies have found little evidence for dominance variance (Ober, Abney et al. 2001; Weiss, Pan et al. 2006; Visscher, Macgregor et al. 2007), although a fourth estimated dominance variance as 9% for men and 7% for women (Eaves, Martin et al. 1999). Absence of evidence for dominance variance need not, however, be inconsistent with evidence of inbreeding depression: it can be shown that with a large number of loci it is theoretically possible to have inbreeding depression in the absence of evidence for dominance variance (*personal communication*, P Visscher).

The results of the present study are, then, consistent both with published inbreeding studies and with genetic theory, given a very large number of loci influencing height. It is likely, then, that the observed association between both H_{pn} and F_{ROH} and height is a real, albeit small, one. This study is not, however, without its drawbacks. Weiss et al (2006) recommend that heritability be estimated for males and females separately because of the possibility of sexual dimorphism. It was decided not to do

this in the present case for the reasons explained in section 5.4.2 above, although it is recognised that this will not fully account for any effects of sexual dimorphism.

Age is another issue warranting further investigation. It is known that old people shrink, due to osteoporosis, etc, thus differences in age profile between samples may affect results, if for example one sample has a higher proportion of old people than another. Again, adjusting for age might not deal with this effect fully. One approach would be to restrict the analysis to a given age band, although the disadvantage of this approach would be a reduction in sample sizes. An alternative approach would be to use demispan (the length from sternal notch to fingertip). This is very highly correlated with adult height and does not change during the aging process, however it was not measured in the cohorts available.

Finally, the highly divergent results in the NSPHS sample remain puzzling. Ascertainment bias, resulting from the fact that markers chosen for polymorphism in other populations, are being used here, may be a factor. This may account for the relatively low h^2 estimate for height in NSPHS (although this estimate was not statistically different from h^2 estimates for the other populations).

5.4.5 Other traits

There was no significant evidence of inbreeding depression for any of the other traits examined, although respiratory traits may warrant further investigation in bigger samples (data were only available for CROAS and ORCADES), particularly as Campbell et al found evidence of inbreeding depression in a related lung function

test (FEF25) (Campbell, Carothers et al. 2007) and Weiss and colleagues were able to estimate dominance variance for FEV1 in Hutterites (2006). Only CROAS and ORCADES provided data on FVC: confidence intervals (table 5.11) suggest that a larger sample size might yield significant results.

5.5 Conclusions

This study investigated recessive effects on 11 QT. The effects of recent inbreeding were investigated using $F_{ROH1.5}$ and F_{ROH5} and the effects of total inbreeding, including that of very ancient origin, were investigated using H_{pn}. The study found no evidence of inbreeding depression on lipid or hypertension-related traits, although failure to see an effect may be in part to do with the difficulties inherent in investigating such QT in populations with high levels of treatment for hypertension and raised cholesterol. The only QT to exhibit evidence of inbreeding depression was height, where a 1% increase in F_{ROH} was associated with a reduction in height of between 0.25 and 0.5 cm. After adjustment for multiple testing, results were significant for both H_{pn} and F_{ROH5}. However, results for one population (NSPHS) were so much higher than the other samples as to give cause for concern. Consequently, data were reanalysed without this population, and whilst results remained nominally significant at the 0.05 level, they were no longer significant after adjustment for multiple testing. There is considerable evidence from published inbreeding studies to support the hypothesis of an inbreeding effect on height and this is theoretically possible despite the lack of evidence of dominance variance for height. For this reason, further investigation of this question using much larger sample sizes would be of interest. P-values for height were stronger for H_{pn} than for

either F_{ROH} measure, which focus on the effects of recent inbreeding. Results for the effect of H_{pn} controlled for inbreeding (i.e. isolating the effects of ancient inbreeding) were non-significant. This suggests that it is the cumulative effect of all homozygosity, not just the homozygosity resulting from recent inbreeding, that is important.

Chapter 6: Homozygosity and colorectal cancer

6.1 Introduction

The previous chapter explored the question of whether recessive effects could be detected in a range of QT, primarily in populations with high levels of parental relatedness compared with urban European populations. F_{ROH} (expressed as a percentage) was used to estimate the effects of recent and deeper inbreeding, whilst H_{pn} was used to estimate the total effects of inbreeding even from ancient ancestry. This final chapter extends the analysis to two samples from much more cosmopolitan UK populations, in order to investigate recessive effects in colorectal cancer risk. Again, both H_{pn} (an estimate of total homozygosity, or both ancient and recent parental relatedness) and two F_{ROH} measures are used: the percentage of the typed autosomal genome in ROH < 1 Mb ($F_{ROH<1}$) is a measure of parental relatedness of very ancient origin, or equivalently homozygosity resulting from small population size; the percentage of the typed autosomal genome in ROH ≥ 1 Mb ($F_{ROH\geq1}$) is a measure of recent parental relatedness.

Although twin studies estimate that genetic susceptibility accounts for around 35% of CRC aetiology in populations of European origin (Lichtenstein, Holm et al. 2000), most of this risk remains unexplained. Some of this "missing heritability" may be attributable to CNV, which current arrays do not cover well and which consequently have not been studied properly. Only a small proportion of CRC heritability can be explained by currently identified susceptibility loci (Houlston, Webb et al. 2008). GWAS have high power to detect common variants (i.e. variants with a population frequency of at least 10 – 20%) with a large influence on risk (i.e. explaining at least

1% of inherited risk), so it is unlikely that there are many more common risk alleles with large effect sizes to be identified in populations of European origin (Houlston, Webb et al. 2008; Tomlinson, Webb et al. 2008). The proportion of heritability explained by the variants identified to date is, however, very low, accounting for less than 5% of inherited risk (Houlston, Webb et al. 2008). This implies that much of the remaining variation in CRC risk is explained by a polygenic model involving many common variants of individually very small effect size, and rarer variants of both small and large effect size, all of which GWAS have low power to detect (Houlston, Webb et al. 2008). Evidence of an association of F_{ROH} and/or H_{pn} on CRC risk would be consistent with this polygenic model and would support the hypothesis that CRC risk is influenced by the combined effect of many recessive variants spread throughout the genome. Here, this hypothesis is investigated using data from two British CRC case-control data sets.

6.2 Methods

6.2.1 Sample details

The London sample consists of 618 cases with colorectal neoplasia and at least one 1st degree relative affected by CRC. The 963 controls are individuals unaffected by cancer and with no family history (up to 2nd degree relatives) of colorectal neoplasia. All are of European ancestry and from the UK (Houlston, Webb et al. 2008). The Scottish sample is from a prospective population-based study conducted in Scotland from 1999 (SOCCS), comprising 980 cases with a confirmed diagnosis of adenocarcinoma of the large bowel. The sample is enriched for early onset cases (diagnosed at age 55 or younger). The 1002 controls are unrelated individuals not

affected by cancer and matched to cases by age (within 5 years), sex and area of residence in Scotland (Tenesa, Farrington et al. 2008).

The London sample was genotyped with the Illumina HumanHap550 BeadChip array, comprising 555,352 SNPs. After performing QC procedures (Houlston, Webb et al. 2008) 547,487 SNPs remained. The SOCCS sample was genotyped with the Illumina HumanHap300 and HumanHap240S arrays, comprising 555,510 SNPs. After performing QC procedures (Houlston, Webb et al. 2008) 548,586 SNPs remained.

Further QC procedures were then performed on both samples. Individuals with more than 5% missing genotypes were excluded. SNPs were excluded if more than 10% were missing or if they failed HWE at $p < 0.0001$. Final sample numbers are shown in table 6.1. Five individuals were removed from the SOCCS sample for low genotyping and 487 SNPs were removed for failing HWE. No individuals were removed from the London sample for low genotyping. 970 SNPs were removed for failing HWE.

A consensus panel was then made, including 540,565 SNPs that satisfied these QC criteria in both samples. 525,727 of these were autosomal SNPs. Excluding the centromeres, the length of the autosomal genome covered by this panel is 2673.83 Mb. This gives a mean density of SNP coverage of 5.23 kb/SNP.

Table 6.1: Sample details

Sample	Male	Female	Total
SOCCS cases	495	481	976
SOCCS controls	513	488	1001
SOCCS total	**1008**	**969**	**1977**
London cases	279	339	618
London controls	439	524	963
London total	**718**	**863**	**1581**
Total cases for meta analysis	**774**	**820**	**1594**
Total controls for meta analysis	**952**	**1012**	**1964**
Total for meta analysis	**1726**	**1832**	**3558**

6.2.2 Definition of F_{ROH}

F_{ROH} statistics were derived using the Runs of Homozygosity programme implemented in PLINK, as described in detail in chapter 3. The following parameters were used:

- The minimum number of consecutive homozygous genotypes constituting a ROH was set at 25
- The minimum length of ROH was set at 150 kb.
- The maximum density (kb/SNP) of a ROH was set at 20
- The maximum gap (kb) between 2 consecutive homozygous SNPs in a ROH was set at 100.
- All other parameters used PLINK defaults (Purcell 2007).

The lengths of all ROH < 1 Mb and all ROH ≥ 1 Mb were summed for each individual and expressed as a percentage of the typed autosomal genome. Case and

control group means were calculated for each sample and the differences between groups were tested for significance. This was done initially for males and females separately within each sample to check for effect modification by sex.

6.2.3 Definition of H_{pn}

H_{pn} was derived from the heterozygosity programme in PLINK, by subtracting the number of homozygous genotypes per individual from the total number of typed SNPs per individual. Case and control group means were calculated for each sample and the difference between groups was tested for significance. This was done initially for males and females separately within each sample to check for effect modification by sex.

6.2.4 Association of CRC with recent inbreeding and distant shared ancestry

In order to understand the extent to which CRC risk is associated with recent inbreeding and the extent to which it is associated with distant shared parental ancestry, the samples were sub-divided into inbred and outbred groups. The outbred group was derived using parameters from the half Orcadian group described in chapter 3. This group consists of individuals with one set of Orcadian and one set of mainland Scottish-born grandparents, with no pedigree evidence of inbreeding in the previous 4 – 5 ancestral generations and a high probability (because of what is known about migration patterns between Orkney and the rest of Scotland) of no inbreeding in about 10 ancestral generations. The maximum value of $F_{ROH1.5}$ in this group (estimated using the Hap300 panel) was 0.52%. $F_{ROH1.5}$ statistics were estimated for the SOCCS and London samples, also using the Hap300 panel, and

subjects with $F_{ROH1.5}$ lower than 0.52% were classified as outbred. This inbreeding threshold is very low, which is appropriate for defining an outbred group: if this were an epidemiological test for inbreeding, it would have very high specificity (i.e. it correctly identifies a high proportion of non-inbred subjects). A much more sensitive definition of inbreeding is, however, required to minimise the risk of false positives in the inbred group. Using data on the offspring of first and second cousins in the ORCADES study described in chapter 3, inbred subjects were defined as those with at least 0.75% of their typed autosomal genome in ROH $\geq$ 1.5 Mb, as measured using the Hap300 panel.

Two analyses were performed. Firstly, in order to investigate whether recessive effects due to distant shared ancestry can be detected, the outbred group was analysed separately. Mean differences between cases and controls were analysed using two measures: H_{pn}, which estimates total homozygosity, and $F_{ROH<1}$, which estimates the effects of non-recent inbreeding. As suggested in chapter 4, this latter measure should be treated with a degree of caution due to the limitations of the Hap500 SNP panel in detecting short ROH.

Secondly, in order to assess the association between CRC risk and recent inbreeding, CRC odds ratios (OR) were estimated for the inbred compared with the outbred groups. ORs were first estimated in groups stratified by sample and sex. A pooled OR was then estimated using Mantel-Haenszel methods (Kirkwood and Sterne 2003).

6.2.5 Meta-analysis

The results from both samples were combined in a meta-analysis (Whitehead 2002). Firstly, the absolute differences between case and control means in each sample were calculated as follows:

Table 6.2: Meta-analysis notation

Data	Cases	Controls	Total
Number of independent studies			r
Number of subjects	n_T	n_C	n
Mean	$\bar{y}_T$	$\bar{y}_C$	
Standard deviation	s_T	s_C	
Sum of observations	A_T	A_C	A
Sum of squares of observations	B_T	B_C	B

The difference between the sample means is:

$$\hat{\theta} = \bar{y}_T - \bar{y}_C$$

This has variance:

$$\text{var}(\hat{\theta}) = \sigma^2 \left(\frac{1}{n_T} + \frac{1}{n_C} \right)$$

σ^2 is estimated using the sample standard deviation of each total sample (i.e. the standard deviation of the cases and controls in each study):

$$s^2 = \frac{B_T - A_T^2 / n_T + B_C - A_C^2 / n_C}{n - 2}$$

Next, results of the two studies were combined. The pooled variance (i.e. pooled between both studies) s^2_{pooled} is derived from the variance of the separate studies above:

$$s^2_{pooled} = \frac{\sum_{i=1}^{r}(n_i - 2)s_i^2}{\sum_{i=1}^{r}(n_i - 2)}$$

where n_i is the total number of subjects from study i.

The variance is derived from s^2_{pooled} using the same equation as above, except that σ^2 is replaced by s^2_{pooled}:

$$\text{var}(\hat{\theta}) = s^2_{pooled}\left(\frac{1}{n_T} + \frac{1}{n_C}\right)$$

The null hypothesis is that the difference in both studies = 0. This is tested by comparing the U statistic with the χ^2 distribution with 1 df.

Let w_i be the estimated inverse variance of $\hat{\theta}_i$:

$$w_i = \frac{1}{\text{var}(\hat{\theta}_i)}$$

then:

$$U = \frac{\left(\sum_{i=1}^{r}\hat{\theta}_i w_i\right)^2}{\sum_{i=1}^{r} w_i}$$

6.2.6 Adjustment for multiple testing

A Bonferroni correction was applied to take account of multiple testing. With an alpha level of 0.05 and 5 tests, the adjusted p-value is 0.01.

6.3 Results

6.3.1 Differences between the SOCCS and London samples

Population distributions for $F_{ROH<1}$, $F_{ROH\geq1}$ and H_{pn} are shown in figures 6.1 and 6.3, separately for cases and controls in each. SOCCS cases had significantly greater $F_{ROH<1}$ (a measure of distant parental relatedness, or small N_e) than did London cases ($p = 0.002$). This effect was even stronger when SOCCS and London controls were compared ($p = 0.00001$). There was no significant difference between SOCCS and London for either cases or controls for $F_{ROH\geq1}$ (a measure of more recent inbreeding) or for H_{pn} (a measure of total homozygosity, or the effects of all inbreeding and shared ancestry, both recent and ancient). Data are shown in table 6.3.

Table 6.3: Comparison of homozygosity statistics between SOCCS and London samples, split by case status (F_{ROH} statistics are expressed as a percentage of the typed autosomal genome)

Group	Statistic	$F_{ROH<1}$	$F_{ROH\geq1}$	H_{pn}
Cases	**Mean difference (SOCCS - London)**	0.073	0.032	0.0001
	SE	0.024	0.028	0.00016
	p	0.002	ns	ns
Controls	**Mean difference (SOCCS - London)**	0.092	0.023	0.0002
	SE	0.021	0.017	0.00012
	p	0.00001	ns	ns

6.3.2 Differences between cases and controls

In all sub-groups, regardless of sample, sex, or homozygosity measure used, cases were more homozygous than controls. Because of this consistency of the direction of effect across sex groups, further analyses were therefore performed on total samples, not split by sex. In both samples cases had on average higher $F_{ROH<1}$,

$F_{ROH\geq1}$ and H_{pn} than did controls. This difference was nominally significant in the meta-analysis for $F_{ROH<1}$ ($p = 0.02$) and remained significant after adjustment for multiple testing for H_{pn} ($p = 0.0085$). Results are shown in table 6.4. F_{ROH} data are shown graphically in figures 6.4 – 6.6.

Table 6.4: Mean difference (cases – controls) in $F_{ROH<1}$, $F_{ROH\geq1}$ and H_{pn} (F_{ROH} statistics are expressed as a percentage of the typed autosomal genome)

Sample	Statistic	$F_{ROH<1}$	$F_{ROH\geq1}$	H_{pn}
SOCCS	**Mean difference**	0.027	0.020	0.0002
	SE	0.017	0.018	0.0001
	p	Ns	Ns	0.051
London	**Mean difference**	0.046	0.011	0.0003
	SE	0.027	0.028	0.00017
	p	Ns	Ns	Ns
Meta-analysis	**Mean difference**	0.035	0.016	0.00024
	SE	0.015	0.015	9.2 E-05
	p	0.02	Ns	0.0085

Figure 6.1: Population distributions for H_{pn}

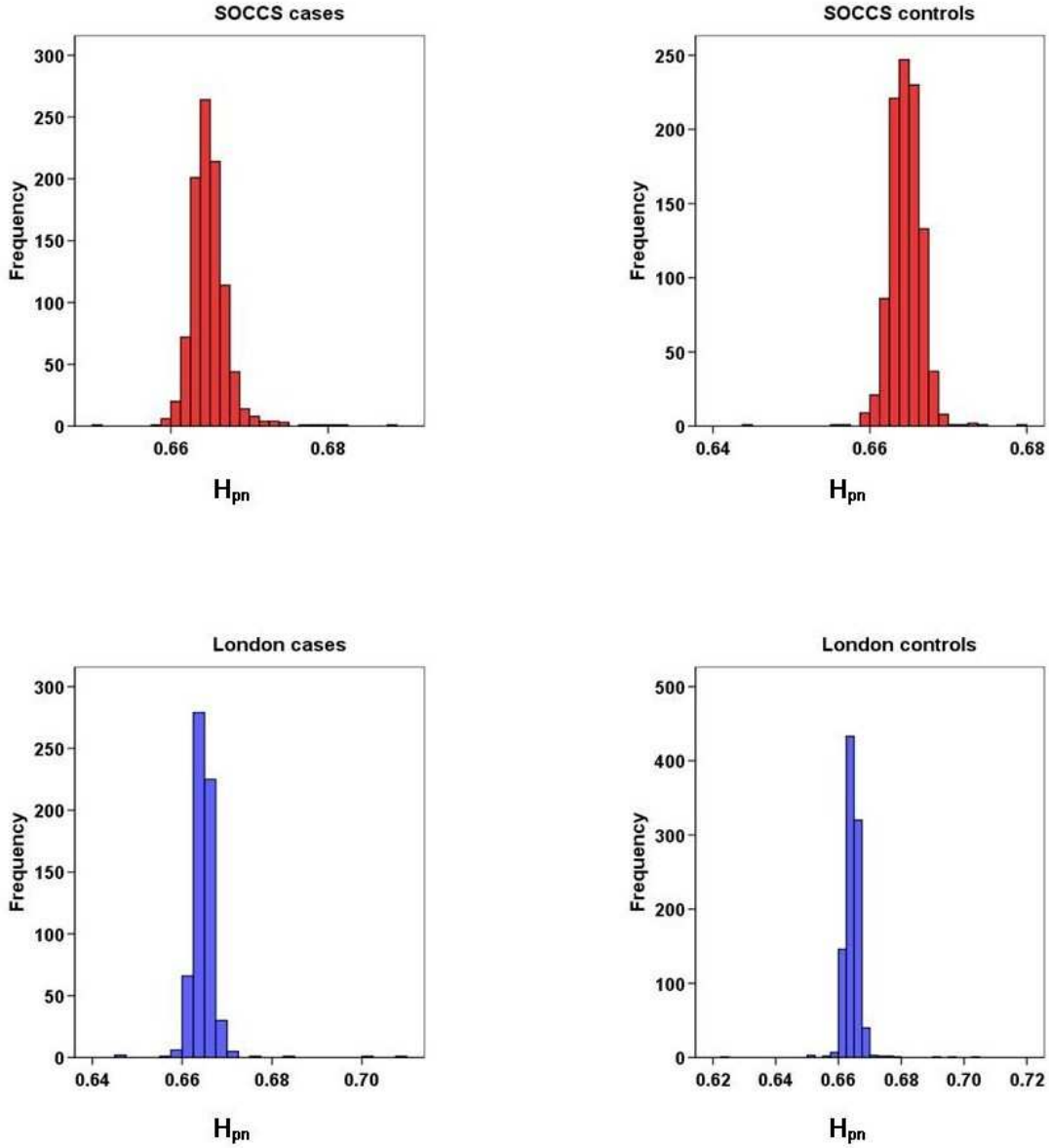


Figure 6.2: Population distributions for $F_{ROH<1}$ (expressed as a percentage)

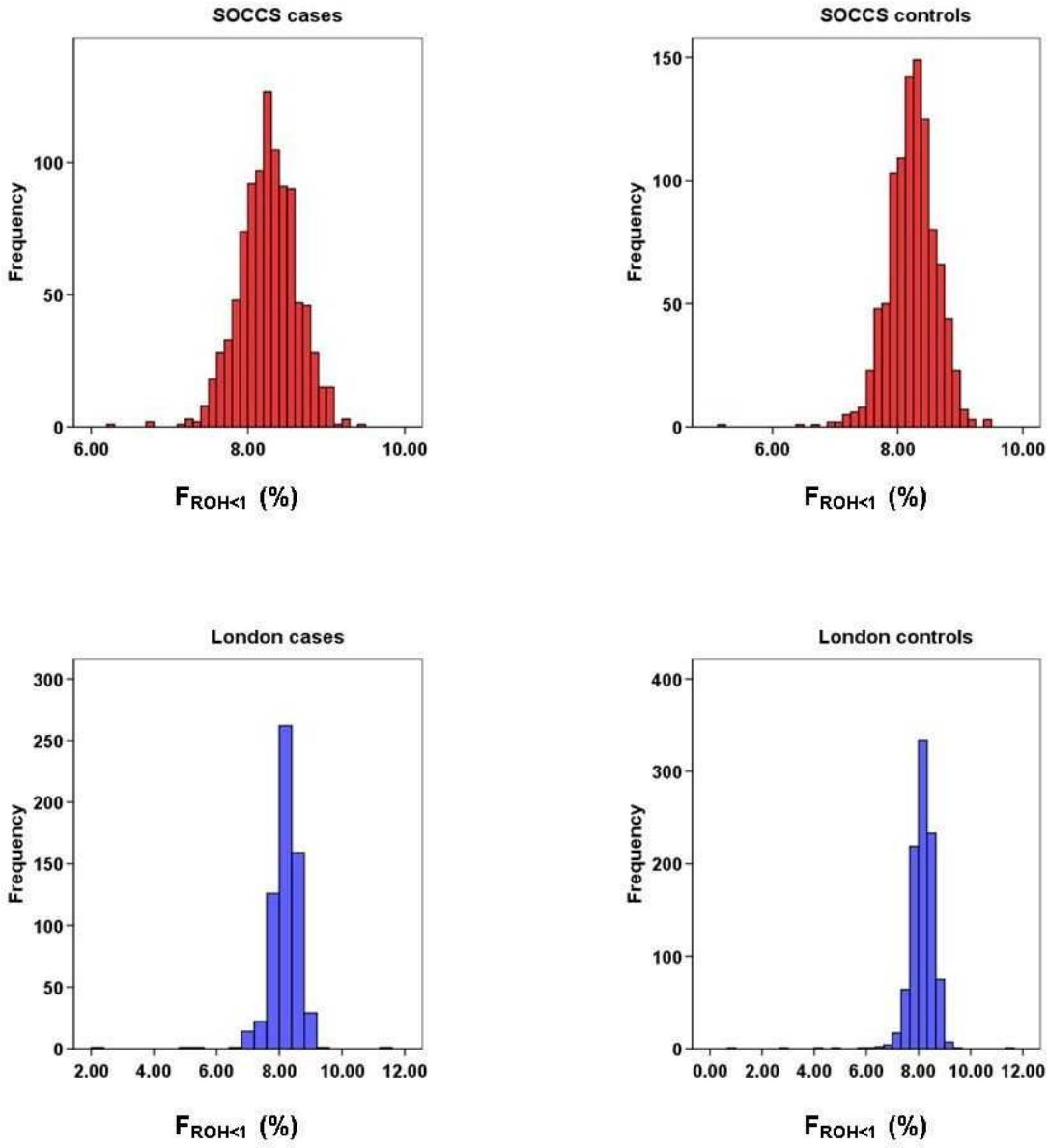


Figure 6.3: Population distributions for $F_{ROH \geq 1}$ (expressed as a percentage)

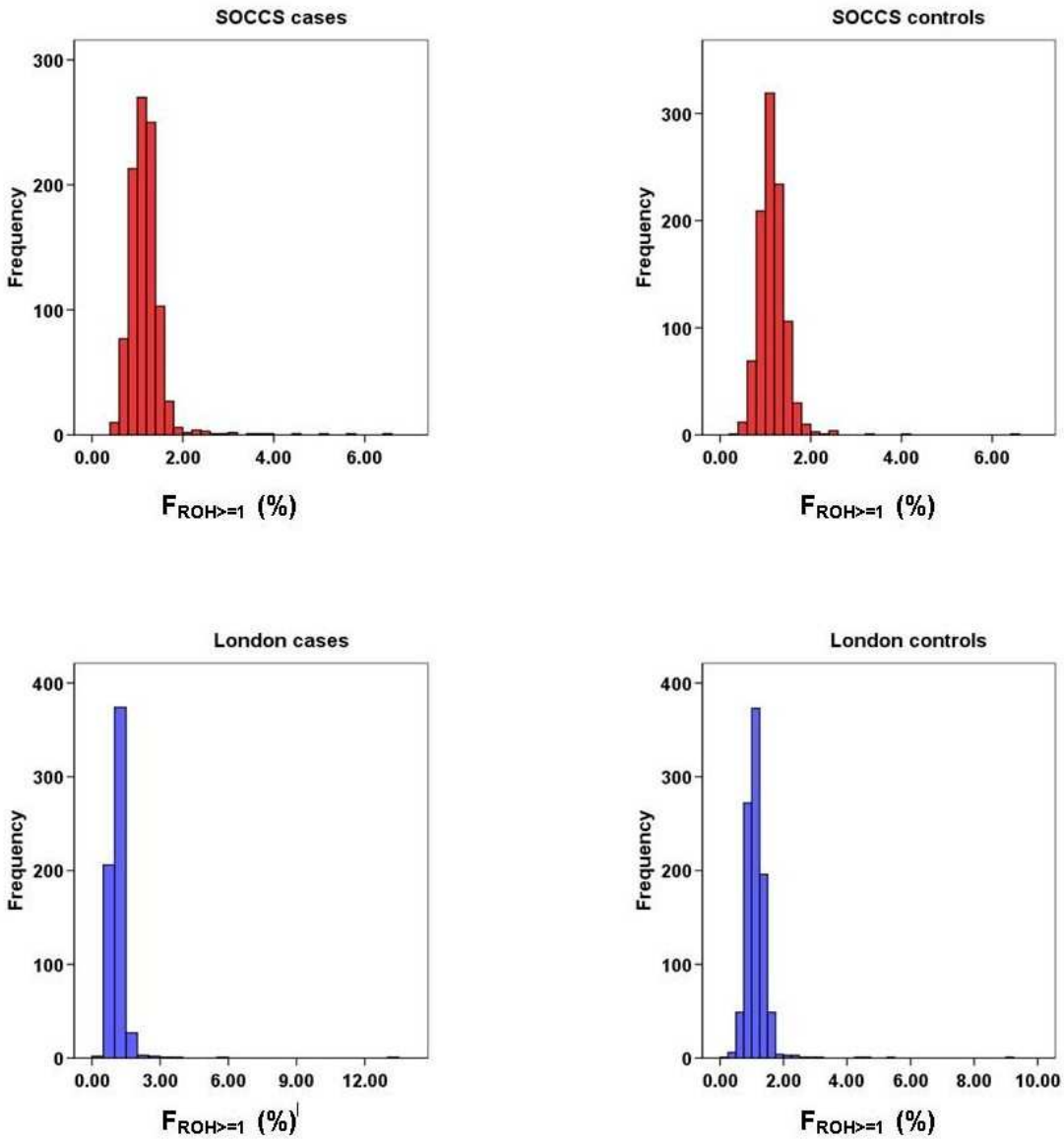


Figure 6.4: Mean and 95% confidence interval for the difference (cases – controls) in percentage of typed autosomal genome in ROH for ROH longer and shorter than 1 Mb (SOCCS)

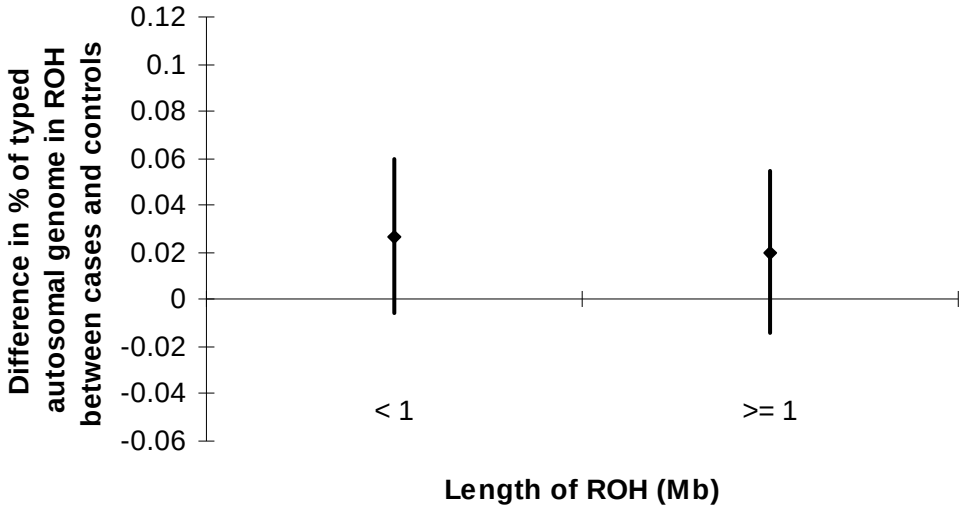


Figure 6.5: Mean and 95% confidence interval for the difference (cases – controls) in percentage of typed autosomal genome in ROH for ROH longer and shorter than 1 Mb (London)

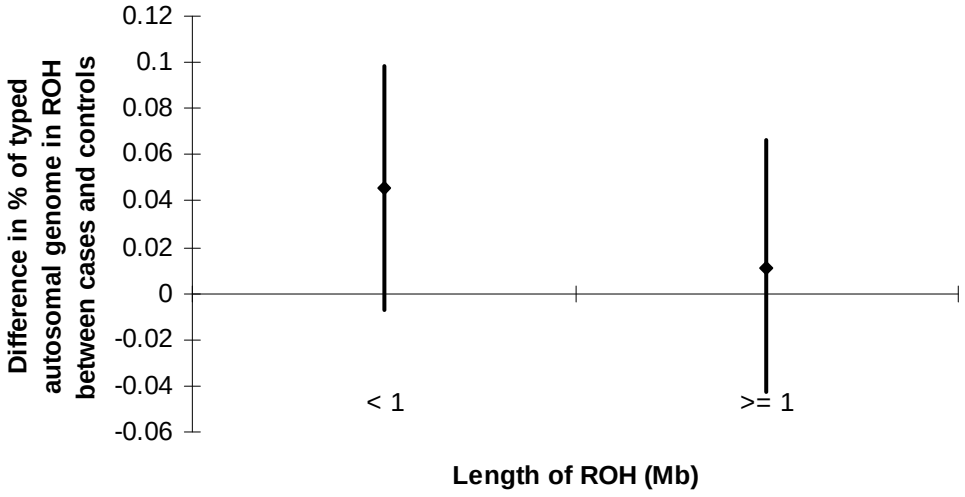
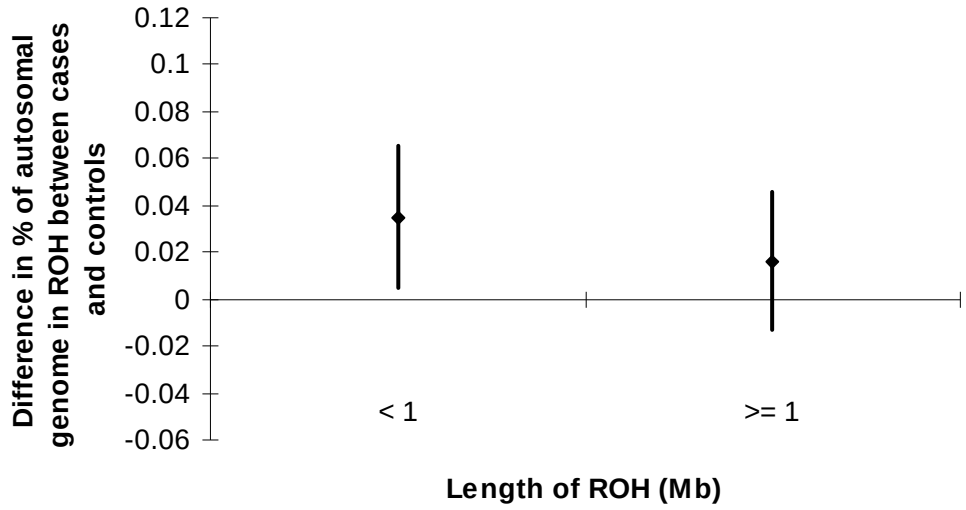


Figure 6.6: Mean and 95% confidence interval for the difference (cases – controls) in % of typed autosomal genome in ROH for ROH longer and shorter than 1 Mb (meta-analysis)



6.3.3 The effects of inbreeding and distant shared ancestry

Finally, two analyses were done to investigate, firstly the extent to which the excess homozygosity observed in cases is due to recent inbreeding and secondly, whether inbred individuals have higher odds of CRC than outbred individuals. Using the approach described in the Methods section above, the samples were partitioned into inbred and non-inbred subjects (table 6.5).

Table 6.5: Inbred and Non-inbred subjects in the SOCCS and London samples

Sample	Sex	Category	Cases	Controls	Total
SOCCS	**Male**	**Inbred**	15	11	26
		Outbred	462	483	945
		Total	477	494	971
	Female	**Inbred**	13	7	20
		Outbred	446	456	902
		Total	459	463	922
London	**Male**	**Inbred**	4	10	14
		Outbred	262	417	679
		Total	266	427	693
	Female	**Inbred**	7	7	14
		Outbred	324	501	825
		Total	331	508	839
SOCCS	**Both**	**Inbred**	28	18	46 (2.3%[1])
		Outbred	908	939	1847
		Total	936	957	1893
London	**Both**	**Inbred**	11	17	28 (1.8%[1])
		Outbred	586	918	1504
		Total	597	935	1532
Combined	**Both**	**Inbred**	39	35	74 (2.1%[1])
		Outbred	1494	1857	3351
		Total	1533	1892	3425[2]

[1] Inbred subjects as a percentage of total sample
[2]133 subjects were excluded from this analysis because they did not meet criteria for either outbred or inbred categories – see Methods section.

The outbred sample was analysed separately in order to investigate whether there was any significant difference in homozygosity between cases and controls after controlling for the effects of inbreeding (table 6.6). Two measures were used: H_{pn}

and $F_{ROH<1}$. Cases were more homozygous than controls according to both measures and in both samples. Differences were nominally significant for H_{pn} in the SOCCS sample (p = 0.03), and (just) significant for H_{pn} after correction for multiple testing in the meta analysis (p = 0.01).

Table 6.6: Mean difference between cases and controls in SOCCS and London "outbred" sub-sets (F_{ROH} expressed as percentage)

Sample	Statistic	H_{pn}	$F_{ROH<1}$
SOCCS (n = 1847)	**Mean difference**	0.0002	0.022
	SE	9.4×10^{-5}	0.017
	p	0.03	ns
London (n = 1504)	**Mean difference**	0.0002	0.034
	SE	0.00012	0.027
	p	ns	ns
Meta-analysis (n = 3351)	**Mean difference**	0.0002	0.027
	SE	7.8×10^{-5}	0.015
	p	0.01	ns

Secondly, odds ratios were calculated in order to investigate whether there is any evidence that inbreeding increases the odds of CRC. None of the sub-group OR were significant; however in all groups except London males the odds of CRC were higher in outbred than in inbred subjects. The odds of CRC in inbred compared with outbred subjects were statistically significant in the pooled sample. Results are shown in table 6.7.

Table 6.7: CRC OR for inbred compared with outbred subjects

Sample	OR	95% confidence interval
SOCCS males	1.43	0.65 – 3.14
SOCCS females	1.90	0.75 – 4.80
London males	0.64	0.20 – 2.05
London females	1.55	0.54 – 4.45
Combined	1.36	1.08 – 1.71

6.4 Discussion

Using data from two predominantly outbred UK populations, this study shows that, regardless of how it is measured, homozygosity is consistently higher in CRC cases than controls. This difference was nominally significant in the meta-analysis for $F_{ROH<1}$ (a measure of the effects of ancient inbreeding or small N_e) and significant after adjusting for multiple testing in the meta-analysis for H_{pn} (a measure of total homozygosity). As discussed in chapter 4, the Hap500 panel under-estimates the prevalence of short ROH: for this reason, the true effect of short ROH is likely to be under-estimated.

Although the samples used for this analysis have very low levels of inbreeding (an estimated 1 – 2% of subjects met the genomic criteria for inbreeding) some evidence of increased risk of CRC amongst inbred subjects was found. The odds of CRC were (slightly) significantly higher for inbred than non-inbred subjects. These findings are consistent with the results of a recent study by Bacolod and colleagues, who found that CRC cases were twice as likely to have extended IBD segments than were

controls (Bacolod, Schemmann et al. 2008). For $F_{ROH \geq 1}$ (a measure of recent inbreeding) on average cases had higher values than controls, although differences were not significant. In order to investigate recessive effects resulting from remote common ancestry as opposed to recent inbreeding, the non-inbred category was analysed separately. On average cases still had higher H_{pn} and $F_{ROH<1}$ than did controls and these differences were nominally significant for H_{pn} in the SOCCS sample ($p = 0.03$) and (just) significant after adjustment for multiple testing in the meta-analysis of both samples ($p = 0.01$). Even in outbred samples, levels of homozygosity vary as a consequence of individual ancestral demographic history. This study suggests that such variation, which is beyond the reach of traditional pedigree analysis and can only be detected using genomic measures, may contribute to CRC risk.

These findings are consistent with a polygenic model of CRC aetiology and suggest that the combined effects of many recessive alleles of individually small impact, scattered throughout the genome, exert an influence on CRC risk. The statistical models used by GWAS to date have generally been based on additive effects, with little attempt made to identify recessive effects. The next step in this research is to see whether these results can be replicated in other CRC samples. There are also plans to use this approach to assess the role of polygenic recessivity as a novel risk factor in a range of other complex diseases.

Finally, differences between the SOCCS and London samples are interesting. SOCCS cases and to an even greater extent, controls, had significantly more of their

genome in short ROH compared with London cases and controls. This presumably reflects differences in demographic history and N_e between the two populations from which the samples are drawn, with the SOCCS sample characterised by smaller N_e and higher levels of distant parental relatedness compared with the London sample. Given the suggested association between homozygosity and CRC risk, these population-level differences may be of interest in explaining some of the variability in CRC risk among populations.

Chapter 7: Conclusions

There is a widespread public perception, dating from the nineteenth century, that marriage between kin is detrimental to health. More recently, the connection between genetic disease and consanguinity has been reinforced in popular consciousness by the success of homozygosity mapping in identifying the alleles responsible for rare, recessive monogenic disease in consanguineous families (Bittles 2008). Unfortunately, this emphasis on (often poorly understood) adverse health outcomes is rarely balanced against the undoubted social advantages of consanguineous unions to many, particularly in poorer parts of the world. Such a negative focus can be distressing and stigmatising for consanguineous families and communities (Bittles 2008).

From a genetic perspective, inbreeding is simply a mechanism for increasing homozygosity: it is increased homozygosity for deleterious recessive alleles that is of interest to those seeking to understand the genetic architecture of complex disease. Until quite recently, the only way to quantify individual homozygosity was by reference to inbreeding: i.e. by estimating F_{ped}. The development of dense genome scanning means that today there is the potential to investigate homozygosity and recessive effects directly, without reference to inbreeding or consanguinity. This will enhance the ability of researchers to investigate recessive effects in outbred, as well as in more unusual, populations. It has the added benefit of shifting the focus of

research onto homozygosity and away from inbreeding, with all its stigmatising connotations.

This study set out to develop a genomic measure of homozygosity, F_{ROH}, that would encapsulate the genomic effects of an individual's entire ancestral demographic history. This measure is rooted in meiosis, the fundamental biological process determining chromosomal configuration (figure 1.1). It is based on the fact that homozygous genotypes are not evenly distributed throughout the genome but are distributed in extended tracts, the only truly sporadic homozygous genotype being one that has arisen by mutation. It therefore follows that, excepting the case of homozygous genotypes that have arisen as a result of mutation, the distinction between autozygosity and homozygosity is one of degree, not of substance (although with the caveat that recent parental relatedness is more likely to result in homozygosity for rare alleles, which are more often deleterious). If it were routinely possible to observe the genome directly, quantifying the genomic effects of an individual's shared maternal and paternal ancestry would be a simple matter of summing the length of all his ROH, from the longest to the shortest. Furthermore, the distribution of these ROH by different length categories would provide information about N_e and inbreeding loops at different points in his ancestors' demographic history: in general, the shorter the ROH, the more distant the common ancestor from whom it originates, although there is considerable variation because of the random nature of recombination during meiosis.

In practice, of course, technology imposes constraints on the length of ROH it is possible to detect, and therefore the distance back in time it is possible to look. Chapter 4 addresses this issue in some detail, suggesting limits of reliability for SNP panels commonly used in human studies. Because 500K and 300K panels do not reliably detect ROH shorter than 1 Mb and 1.5 Mb respectively, at present this is not the best approach to quantifying short ROH resulting from very distant shared parental ancestry. Whilst this study found F_{ROH} to be significantly more strongly correlated with F_{ped}, and thus better at estimating the genomic effects of recent parental relatedness, than any of the other measures investigated (chapter 3), H_{pn} (the proportion of typed SNPs that are homozygous) is more effective at estimating the genomic effects of very remote parental relatedness. In future, however, much denser SNP panels will become routinely available, opening up the possibility of a much more observational approach to quantifying and analysing homozygosity. This future is not that far away: 1 million SNP panels are increasingly widely used, the 1000 Genomes project is expected to publish data later this year and it will only be a matter of time before whole genome sequencing on a large scale becomes economically feasible.

But returning to the present, by applying these approaches to QT data in genetically isolated population samples, this study provides evidence suggestive of recessive genetic effects on height (chapter 5). This was unexpected, as little evidence has been published for dominance variance in height, although it is consistent with several published studies reporting a reduction in height associated with inbreeding. Both recent inbreeding and homozygosity of much more distant origin contribute to

the recessive effects observed. Some researchers have speculated that beneficial temporal trends in height, and also in other QT such as age at menarche, IQ and lifespan, might be partially attributable to falling levels of consanguinity and endogamy resulting from increasing population mobility over the last century. There is evidence that individuals are becoming less homozygous with each successive generation: Nalls and colleagues found a significant trend of decreased autozygosity with younger chronological age in an outbred North American cohort (2009). Data from the EUROSPAN populations are consistent with this: age was significantly associated with H_{pn} in all samples (CROAS, $p = 2.6 \times 10^{-8}$; ERF, $p = 0.013$; MICROS, $p = 2.3 \times 10^{-8}$; NSPHS, $p = 1.5 \times 10^{-6}$; ORCADES, $p = 1.6 \times 10^{-7}$) suggesting an ongoing process of isolate breakdown in these populations. On the other hand, there was no association between age and H_{pn} in the SOCCS sample, which more closely resembles that in the Nalls study.

Contrary to several published studies, no evidence was found for recessive effects on blood pressure or cholesterol, although studying these traits in populations with high levels of treatment for hypertension and hypercholesterolaemia is not without its methodological problems, which may have been a factor here.

Applying F_{ROH} and H_{pn} to CRC case-control data in two non-isolate population samples, this study found evidence of recessive effects on CRC risk (chapter 6). Again, effects were attributable both to recent inbreeding and to more distant parental relatedness. Both analyses demonstrate clearly that homozygosity originating from distant shared parental ancestry, sometimes described as

background homozygosity, contributes to recessive effects. Background homozygosity differs among populations, reflecting differences in population history and N_e, and as such, might contribute to differences in disease risk from one population to another. It has nothing to do with inbreeding, as it is commonly understood, or to consanguinity, and is beyond the reach of pedigree-based measures of autozygosity.

Immediate priorities for taking this work forward fall into two broad categories. Firstly, F_{ROH} could be improved by theoretical modelling. Using LD patterns to infer recombination rates between all adjacent SNP pairs and including this information in a model of homozygosity-by-descent would provide richer information on the probable age and population prevalence of individual ROH and would provide greater confidence that observed consecutive homozygous SNPs truly represent ROH (i.e. that the unobserved SNPs between consecutive homozygous SNPs are also homozygous). This would particularly enhance the ability to identify short ROH (Leutenegger, Labalme et al. 2006; Auton, Bryc et al. 2009). Modelling of allele frequency would also be valuable in this respect, as a run of common alleles has a higher probability of occurring by chance (i.e. not representing a true ROH) than does a run of rare alleles.

Secondly, the discovery of an association between CRC and homozygosity suggests that this approach should be applied to a variety of other diseases. With the recent emergence of large GWAS consortia, typically involving in the region of up to 50,000 genotyped subjects, the current scientific environment is ideal for this type of

endeavour. Following on from the CRC analysis, a priority would be to replicate this in a larger CRC sample, as well as repeating the analysis in other cohorts of common cancers, such as lung, breast and prostate. Agreement has already been secured to analyse data from the Wellcome Trust Case Control Consortium (WTCCC) comprising 19,000 subjects and 7 disease cohorts; the Meta-Analysis of Glucose and Insulin Consortium (MAGIC; glucose and insulin); and the European Network for Genetic and Genomic Epidemiology (ENGAGE; urate). Applications have also been made to Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE; blood pressure); Diabetes Genetics Initiative (DIAGRAM; type 2 diabetes); ENGAGE (lipids) and Genomewide Investigation of Anthropometric Measures (GIANT; height and BMI). In addition, discussions are underway to replicate the height analysis in a much larger sample than was possible in the present study.

Although the current project focuses primarily on complex disease and QT, agreement has been secured with the Malaria Genomic Epidemiology Network (MALARIAGEN) to repeat this analysis in their cohort, which includes data on disease severity, survival and age of onset. A recent study by Lyons and colleagues (2009) found evidence suggestive of increased susceptibility to infectious disease associated with inbreeding (the diseases investigated were tuberculosis and hepatitis B).

Finally, the effects of homozygosity on the health of subjects from culturally consanguineous populations is also of interest. The London Life Sciences Population (LOLIPOP) is a study of premature coronary heart disease involving

24,000 UK Indian Asians and Northern Europeans, which would be very amenable to this approach. Given the global prevalence of consanguinity in countries where this is the cultural norm and also in large immigrant communities in western Europe and north America, this is a subject of epidemiological importance in its own right (Bittles 2008), in addition to being of interest to those investigating recessive effects in complex disease aetiology.

Technological and methodological advances over recent years have revolutionised genetic epidemiology. Many common causal disease variants have been identified by GWAS, leading to greater understanding of disease pathways, the first step on the road to more effective treatments. Although less methodological progress has been made in understanding recessive effects in common disease aetiology, the technology now exists to develop this field. It is an exciting time to be involved.

References

Abecasis, G. R., D. Ghosh, et al. (2005). "Linkage disequilibrium: ancient history drives the new genetics." Hum Hered 59(2): 118-24.

Abecasis, G. R., E. Noguchi, et al. (2001). "Extent and distribution of linkage disequilibrium in three genomic regions." Am J Hum Genet 68(1): 191-197.

Abney, M., M. S. McPeek, et al. (2001). "Broad and narrow heritabilities of quantitative traits in a founder population." Am J Hum Genet 68(5): 1302-7.

Amos, W., J. W. Wilmer, et al. (2001). "The influence of parental relatedness on reproductive success." Proc Biol Sci 268(1480): 2021-7.

Anderson, P. (1988). "The Armada and the North Isles." Northern Studies 25: 42-57.

Aulchenko, Y. S. (2008) "GenABEL Tutorial." Volume, DOI:

Aulchenko, Y. S., P. Heutink, et al. (2004). "Linkage disequilibrium in young genetically isolated Dutch population." Eur J Hum Genet 12(7): 527-34.

Aulchenko, Y. S., S. Ripke, et al. (2007). "GenABEL: an R library for genome-wide association analysis." Bioinformatics 23(10): 1294-6.

Auton, A., K. Bryc, et al. (2009). "Global distribution of genomic diversity underscores rich complex history of continental human populations." Genome Res 19(5): 795-803.

Bacolod, M. D., G. S. Schemmann, et al. (2008). "The signatures of autozygosity among patients with colorectal cancer." Cancer Res 68(8): 2610-21.

Badaruddoza (2004). "Inbreeding effects on metrical phenotypes among north Indian children." Collegium Antropologicum 28 Suppl(2): 311-319.

Baigent, C., A. Keech, et al. (2005). "Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins." Lancet 366(9493): 1267-78.

Bener, A., R. Hussain, et al. (2006). "Consanguineous marriages and their effects on common adult diseases: studies from an endogamous population." Medical Principles and Practice 16: 262-267.

Berry, R. J. (1986). The People of Orkney. The People of Orkney. R. a. F. Berry, HN. Kirkwall, The Orkney Press.

Bittles, A. (2001). "Consanguinity and its relevance to clinical genetics." Clin Genet 60(2): 89-98.

Bittles, A. H. (1990). Consanguineous marriage: current global incidence and its relevance to demographic research., Population Studies Center, University of Michigan.

Bittles, A. H. (2001). "Consanguinity/Endogamy Resource." Retrieved 20/05/2009, 2009, from www.consang.net.

Bittles, A. H. (2003). "Consanguineous marriage and childhood health." Dev Med Child Neurol 45(8): 571-6.

Bittles, A. H. (2008). "A community genetics perspective on consanguineous marriage." Community Genet 11(6): 324-30.

Bolino, A., V. Brancolini, et al. (1996). "Localization of a gene responsible for autosomal recessive demyelinating neuropathy with focally folded myelin sheaths to chromosome 11q23 by homozygosity mapping and haplotype sharing." Hum Mol Genet 5(7): 1051-4.

Botstein, D. and N. Risch (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nat Genet 33 Suppl: 228-37.

Bouhouche, A., A. Benomar, et al. (1999). "A locus for an axonal form of autosomal recessive Charcot-Marie-Tooth disease maps to chromosome 1q21.2-q21.3." Am J Hum Genet 65(3): 722-7.

Bowers, E. J. (1986). Length of life in Orkney. The People of Orkney. F. H. Berry RJ. Kirkwall, The Orkney Press.

Boyce, A. J., V. M. L. Holdsworth, et al. (1973). Demographic and genetic studies in the Orkney islands. Genetic Variation in Britain. D. F. Roberts, Sunderland, E. London, Taylor and Francis. 12.

Brennan, E. R. (1981). "Kinship, demographic, social and geographic characteristics of mate choice in Sanday, Orkney Islands, Scotland." Anthropol 55: 129-38.

Brennan, E. R. and B. Dyke (1980). "Assortative mate choice and mating opportunity on Sanday, Orkney islands." Social Biology 27(3): 199-209.

Brennan, E. R., P. W. Leslie, et al. (1982). "Mate choice and genetic structure Sanday, Orkney Islands, Scotland." Human Biology 54(3): 477-89.

Brennan, E. R. and J. H. Relethford (1983). "Temporal variation in the mating structure of Sanday, Orkney Islands, Scotland." Annals of Human Biology 54(3): 477-89.

Broman, K. W. and J. L. Weber (1999). "Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain." Am J Hum Genet 65(6): 1493-500.

Brown, E. S. (1965). "Distribution of the ABO and Rhesus (D) blood groups in the north of Scotland." Heredity 20: 289-303.

Burton, P. R., M. D. Tobin, et al. (2005). "Key concepts in genetic epidemiology." Lancet 366(9489): 941-51.

Campbell, H., A. D. Carothers, et al. (2007). "Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits." Hum Mol Genet 16(2): 233-41.

Carothers, A. D., I. Rudan, et al. (2006). "Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches." Ann Hum Genet 70(Pt 5): 666-76.

Carstairs, V. and R. Morris (1990). "Deprivation and health in Scotland." Health Bull (Edinb) 48(4): 162-75.

Charlesworth, B. and D. Charlesworth (1999). "The genetic basis of inbreeding depression." Genet Res 74: 329-340.

Charlesworth, B. and K. A. Hughes (1996). "Age-specific inbreeding depression and components of genetic variance in relation to the evolution of senescence." Proc Natl Acad Sci U S A 93(12): 6140-5.

Charpentier, M., J. M. Setchell, et al. (2005). "Genetic diversity and reproductive success in mandrills (Mandrillus sphinx)." Proc Natl Acad Sci U S A 102(46): 16723-8.

Colella, S., C. Yau, et al. (2007). "QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data." Nucleic Acids Res 35(6): 2013-25.

Collacott, R. A. (1984). The pattern of hypertensive disease in the North Isles of Orkney, University of Oxford. PhD thesis.

Collins, A., W. Lau, et al. (2004). "Mapping genes for common diseases: the case for genetic (LD) maps." Hum Hered 58(1): 2-9.

Coltman, D. W., J. G. Pilkington, et al. (1999). "Parasite-mediated selection against inbred Soay sheep in a free-living island population." Evolution 53: 1259-1267.

Coull, J. R. (1966). "Population trends and structures on the island of Westray, Orkney." Scottish Studies 10: 69-77.

Cui, J. S., J. L. Hopper, et al. (2003). "Antihypertensive treatments obscure familial contributions to blood pressure variation." Hypertension 41(2): 207-10.

Curtis, D., A. E. Vine, et al. (2008). "Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations." Annals of Human Genetics 72: 261-278.

DeCroo, S., M. I. Kamboh, et al. (1988). "Isoelectric focusing of superoxide dismutase: report of the unique SOD A*2 allele in a US white population." Human Heredity 38: 1-7.

Durbin, R. and D. Altshuler. (2008). "1000 Genomes: A Deep Catalog of Human Genetic Variation." Retrieved 20/5/2009, 2009, from www.1000genomes.org.

Eaves, L. J., N. G. Martin, et al. (1999). Biological and cultural inheritance of stature and attitudes. Personality and Psychopathology. C. Cloninger. Washington DC, American Psychopathological Association: 269-308.

Eldon, B. J., J. Axelsson, et al. (2001). "Cardiovascular risk factors and relatedness in an Icelandic subpopulation." Int J Circumpolar Health 60(4): 499-502.

Falconer, D. S. and T. F. C. Mackay (1996). Introduction to Quantitative Genetics. Harlow, Essex, UK, Longman Group Ltd.

Fisher, R. A. and G. L. Taylor (1940). "Scandinavian influence in Scottish ethnology." Nature 145: 590.

Flaws, M. and G. Lamb (1996). The Orkney Dictionary. Kirkwall, The Orkney Language and Culture Group.

Florez, J. C. (2008). "Clinical review: the genetics of type 2 diabetes: a realistic appraisal in 2008." J Clin Endocrinol Metab 93(12): 4633-42.

Frazer, K. A. (2003). "The International HapMap Project." Nature 426(6968): 789-96.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature 449(7164): 851-61.

Frazer, K. A., S. S. Murray, et al. (2009). "Human genetic variation and its contribution to complex traits." Nat Rev Genet 10(4): 241-51.

Freire-Maia, N. (1983). "Inbreeding studies in Brasilian schoolchildren." Am J Med Genet 16(3): 331-55.

Gibson, J., N. E. Morton, et al. (2006). "Extended tracts of homozygosity in outbred human populations." Hum Mol Genet 15(5): 789-95.

Gschwend, M., O. Levran, et al. (1996). "A locus for Fanconi anemia on 16q determined by homozygosity mapping." Am J Hum Genet 59(2): 377-84.

Halberstein, R. A. (1999). "Blood pressure in the Caribbean." Hum Biol 71(4): 659-84.

HapMap. (2002). "International HapMap Project." Retrieved 20/5/2009, 2009, from www.hapmap.org.

Hartl, D. L. and A. G. Clark (1997). Principles of Population Genetics. Sunderland MA, USA, Sinauer Associates.

Helgason, A., E. Hickey, et al. (2001). "mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry." Am J Hum Genet 68(3): 723-37.

Hofer, T., N. Ray, et al. (2009). "Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection." Ann Hum Genet 73(1): 95-108.

Hoffman, J. I., I. L. Boyd, et al. (2004). "Exploring the relationship between parental relatedness and male reproductive success in the Antarctic fur seal Arctocephalus gazella." Evolution Int J Org Evolution 58(9): 2087-99.

Houlston, R. S., E. Webb, et al. (2008). "Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer." Nat Genet 40(12): 1426-35.

Hurwich, B. J. and N. Nubani (1978). "Blood pressures in a highly inbred community--Abu Ghosh, Israel. 1. Original survey." Isr J Med Sci 14(9): 962-9.

Isaacs, A., F. A. Sayed-Tabatabaei, et al. (2007). "Heritabilities, apolipoprotein E, and effects of inbreeding on plasma lipids in a genetically isolated population: the Erasmus Rucphen Family Study." Eur J Epidemiol 22(2): 99-105.

Ismail, J., T. H. Jafar, et al. (2004). "Risk factors for non-fatal myocardial infarction in young South Asian adults." Heart 90(3): 259-63.

Johansson, A., F. Marroni, et al. (2009). "Common variants in the JAZF1 gene associated with height identified by linkage and genome-wide association analysis." Hum Mol Genet 18(2): 373-80.

Johansson, A., V. Vavruch-Nilsson, et al. (2005). "Linkage disequilibrium between microsatellite markers in the Swedish Sami relative to a worldwide selection of populations." Hum Genet 116(1-2): 105-13.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res 12(6): 996-1006.

Khlat, M. and M. Khoury (1991). "Inbreeding and diseases: demographic, genetic, and epidemiologic perspectives." Epidemiol Rev 13: 28-41.

Kirkwood, B. R. and J. A. C. Sterne (2003). Essential Medical Statistics. Massachussetts, Blackwell Science Ltd.

Krieger, H. (1969). "Inbreeding effects on metrical traits in Northeastern Brazil." Am J Hum Genet 21(6): 537-46.

Krishan, G. (1986). "Effect of parental consanguinity on anthropometric measurements among the Sheikh Sunni Muslim boys of Delhi." Am J Phys Anthropol 70(1): 69-73.

Lander, E. S. (1996). "The new genomics: global views of biology." Science 274(5287): 536-9.

Lander, E. S. and D. Botstein (1987). "Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children." Science 236(4808): 1567-70.

LaRosa, J. C., J. He, et al. (1999). "Effect of statins on risk of coronary disease: a meta-analysis of randomized controlled trials." Jama 282(24): 2340-6.

Law, M. R., N. J. Wald, et al. (2003). "Value of low dose combination treatment with blood pressure lowering drugs: analysis of 354 randomised trials." Bmj 326(7404): 1427.

Leal, A., B. Morera, et al. (2001). "A second locus for an axonal form of autosomal recessive Charcot-Marie-Tooth disease maps to chromosome 19q13.3." Am J Hum Genet 68(1): 269-74.

Lebel, R. R. and W. B. Gallagher (1989). "Wisconsin consanguinity studies. II: Familial adenocarcinomatosis." Am J Med Genet 33(1): 1-6.

LeGuern, E., A. Guilbot, et al. (1996). "Homozygosity mapping of an autosomal recessive form of demyelinating Charcot-Marie-Tooth disease to chromosome 5q23-q33." Hum Mol Genet 5(10): 1685-8.

Lencz, T., C. Lambert, et al. (2007). "Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia." Proc Natl Acad Sci U S A 104(50): 19942-7.

Leutenegger, A. L., A. Labalme, et al. (2006). "Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome." Am J Hum Genet 79(1): 62-6.

Leutenegger, A. L., B. Prum, et al. (2003). "Estimation of the inbreeding coefficient through use of genomic data." Am J Hum Genet 73(3): 516-23.

Li, L. H., S. F. Ho, et al. (2006). "Long contiguous stretches of homozygosity in the human genome." Hum Mutat 27(11): 1115-21.

Lichtenstein, P., N. V. Holm, et al. (2000). "Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland." N Engl J Med 343(2): 78-85.

Liu, F., A. Arias-Vasquez, et al. (2007). "A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population." Am J Hum Genet 81(1): 17-31.

Liu, F., S. Elefante, et al. (2006). "Ignoring distant genealogic loops leads to false-positives in homozygosity mapping." Ann Hum Genet 70(Pt 6): 965-70.

Lyons, E. J., A. J. Frodsham, et al. (2009). "Consanguinity and susceptibility to infectious diseases in humans." Biol Lett.

Mackenbach, J. P. (1992). "Socio-economic health differences in The Netherlands: a review of recent empirical findings." Soc Sci Med 34(3): 213-26.

Marioni, J. C., N. P. Thorne, et al. (2007). "Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization." Genome Biol 8(10): R228.

Martin, A. O., T. W. Kurczynski, et al. (1973). "Familial studies of medical and anthropometric variables in a human isolate." Am J Hum Genet 25(6): 581-93.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nat Rev Genet 9(5): 356-69.

McLoone, P. (2001). "Carstairs scores for Scottish postcode sectors from the 2001 census." Retrieved 23 April, 2009, from http://www.sphsu.mrc.ac.uk/sitepage.php?page=carstairs.

Miano, M. G., S. G. Jacobson, et al. (2000). "Pitfalls in homozygosity mapping." Am J Hum Genet 67(5): 1348-51.

Miller, R. (1986). Who are the Orcadians? The People of Orkney. F. H. Berry RJ. Kirkwall, The Orkney Press.

Modell, B. and A. Darr (2002). "Science and society: genetic counselling and customary consanguineous marriage." Nat Rev Genet 3(3): 225-9.

Morton, N. E. (1958). "Empirical risks in consanguineous marriages: birth weight, gestation time, and measurements of infants." Am J Hum Genet 10(3): 344-9.

Nalls, M. A., J. Simon-Sanchez, et al. (2009). "Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics." PLoS Genet 5(3): e1000415.

Neel, J. V., W. J. Schull, et al. (1970). "The effects of parental consanguinity and inbreeding in Hirado, Japan. II. Physical development, tapping rate, blood pressure, intelligence quotient, and school performance." Am J Hum Genet 22(3): 263-86.

Ober, C., M. Abney, et al. (2001). "The genetic dissection of complex traits in a founder population." Am J Hum Genet 69(5): 1068-79.

Pardo, L. M., I. MacKay, et al. (2005). "The effect of genetic drift in a young genetically isolated population." Ann Hum Genet 69(Pt 3): 288-95.

Pattaro, C., F. Marroni, et al. (2007). "The genetic study of three population microisolates in South Tyrol (MICROS): study design and epidemiological perspectives." BMC Med Genet 8: 29.

Penrith, J. and D. Penrith (2002). Scottish Islands: Orkney and Shetland. Oxford, Vacation Work.

Purcell, S. (2007). "PLINK v1.0 whole genome association analysis toolset." 1.0. Retrieved 28/6/2008, 2008, from http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet 81(3): 559-75.

Puzyrev, V. P., S. V. Lemza, et al. (1992). "Influence of genetic and demographic factors on etiology and pathogenesis of chronic disease in north Siberian aborigines." Arctic Med Res 51(3): 136-42.

Reich, D. E., M. Cargill, et al. (2001). "Linkage disequilibrium in the human genome." Nature 411(6834): 199-204.

Relethford, J. H. and E. R. Brennan (1982). "Temporal trends in isolation by distance on Sanday, Orkney Islands." Human Biology 54(2): 315-27.

Ritchie, A. (1993). Viking Scotland. London, Historic Scotland.

Roberts, D. F. (1985). "Genetic Structure in Orkney." Man (NS) 20: 131-41.
Roberts, D. F. (1986). "Who are the Orcadians?" Anthrop. Anz. 44(2): 93-104.
Roberts, D. F., M. J. Roberts, et al. (1979). "Inbreeding levels in Orkney islanders." J Biosoc Sci 11: 391-5.
Rogers, T., D. Chandler, et al. (2000). "A novel locus for autosomal recessive peripheral neuropathy in the EGR2 region on 10q23." Am J Hum Genet 67(3): 664-71.
Rudan, I. (1999). "Inbreeding and cancer incidence in human isolates." Hum Biol 71(2): 173-87.
Rudan, I., Z. Biloglav, et al. (2006). "Effects of inbreeding, endogamy, genetic admixture, and outbreeding on human health: a (1001 Dalmatians) study." Croat Med J 47(4): 601-10.
Rudan, I., D. Rudan, et al. (2003). "Inbreeding and risk of late onset complex disease." J Med Genet 40(12): 925-32.
Rudan, I., N. Smolej-Narancic, et al. (2003). "Inbreeding and the genetic complexity of human hypertension." Genetics 163(3): 1011-21.
Saar, K., D. Schindler, et al. (1998). "Localisation of a Fanconi anaemia gene to chromosome 9p." Eur J Hum Genet 6(5): 501-8.
Saleh, E. A., A. A. Mahfouz, et al. (2000). "Hypertension and its determinants among primary-school children in Kuwait: an epidemiological study." East Mediterranean Health Journal 6: 333-337.
SE (2004). Scottish Executive Urban Rural Classification 2003-2004. E. a. R. A. Department.
Shami, S. A., R. Qaisar, et al. (1991). "Consanguinity and adult morbidity in Pakistan." Lancet 338(8772): 954.
Simon-Sanchez, J., S. Scholz, et al. (2007). "Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals." Hum Mol Genet 16(1): 1-14.
Simpson, J. L., A. O. Martin, et al. (1981). "Cancers of the breast and female genital system: search for recessive genetic factors through analysis of human isolate." Am J Obstet Gynecol 141(6): 629-36.
Smith, B. (2001). "The Picts and the Martyrs or did the Vikings kill the native populations of Orkney and Shetland?" Northern Studies 36: 7-32.
Soyannwo, M. A., N. Y. Kurashi, et al. (1998). "Blood pressure pattern in Saudi population of Gassim." Afr J Med Med Sci 27(1-2): 107-16.
Stam, P. (1980). "The distribution of the fraction of the genome identical by descent in finite random mating populations." Genet Res 35: 131-155.
Tenesa, A., S. M. Farrington, et al. (2008). "Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21." Nat Genet 40(5): 631-7.
Tobin, M. D., N. A. Sheehan, et al. (2005). "Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure." Stat Med 24(19): 2911-35.
Tomlinson, I. P., E. Webb, et al. (2008). "A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3." Nat Genet 40(5): 623-30.

Turnbull, F., B. Neal, et al. (2008). "Effects of different regimens to lower blood pressure on major cardiovascular events in older and younger adults: meta-analysis of randomised trials." Bmj 336(7653): 1121-3.

Visscher, P. M., S. Macgregor, et al. (2007). "Genome partitioning of genetic variation for height from 11,214 sibling pairs." Am J Hum Genet 81(5): 1104-10.

Vitart, V., A. D. Carothers, et al. (2005). "Increased level of linkage disequilibrium in rural compared with urban communities: a factor to consider in association-study design." Am J Hum Genet 76(5): 763-72.

Vitart, V., I. Rudan, et al. (2008). "SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout." Nat Genet 40(4): 437-42.

Waisfisz, Q., K. Saar, et al. (1999). "The Fanconi anemia group E gene, FANCE, maps to chromosome 6p." Am J Hum Genet 64(5): 1400-5.

Wall, J. D. and J. K. Pritchard (2003). "Haplotype blocks and linkage disequilibrium in the human genome." Nat Rev Genet 4(8): 587-97.

Weedon, M. N., H. Lango, et al. (2008). "Genome-wide association analysis identifies 20 loci that influence adult height." Nat Genet 40(5): 575-83.

Weiss, L. A., L. Pan, et al. (2006). "The sex-specific genetic architecture of quantitative traits in humans." Nat Genet 38(2): 218-22.

Welch, S. G. (1973). A local Orkney polymorphism - erythrocyte indophenol oxidase frequencies. Genetic Variation in Britain. D. Roberts, Sunderland, E, Symposia of the Society for the Study of Human Biology. 12.

Welch, S. G., J. V. Barry, et al. (1973). "A survey of blood group, serum protein and red cell enzyme polymorphisms in the Orkney Islands." Human Heredity 23: 230-40.

Welch, S. G. and G. W. Mears (1972). "Genetic variants of human indophenol oxidase in the Westray island of the Orkneys." Human Heredity 22: 38-41.

Whitehead, A. (2002). Meta-analysis of controlled clinical trials. Chichester, West Sussex, John Wiley and Sons Ltd.

WHO (1985). Community approaches to the control of hereditary diseases: Report of a World Health Organisation Advisory Group WHO/HGN/WG/85.10

Wilson, J. F., D. A. Weiss, et al. (2001). "Genetic evidence for different male and female roles during cultural transitions in the British Isles." Proc Natl Acad Sci U S A 98(9): 5078-83.

Woods, C. G., E. M. Valente, et al. (2004). "A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR." J Med Genet 41(8): e101.

Wright, A., B. Charlesworth, et al. (2003). "A polygenic basis for late-onset disease." Trends Genet 19(2): 97-106.

Wright, A. F., A. D. Carothers, et al. (1999). "Population choice in mapping genes for complex diseases." Nat Genet 23(4): 397-404.

Wright, S. (1922). "Coefficients of inbreeding and relationship." The American Naturalist 56(645): 330-338.

Zottarelli, L. K., T. S. Sunil, et al. (2007). "Influence of parental and socioeconomic factors on stunting in children under 5 years in Egypt." East Mediterr Health J 13(6): 1330-42.

Appendices

Directorate of Operations

Board Headquarters
Garden House
New Scapa Road
Kirkwall
Orkney KW15 1BQ
www.show.scot.nhs.uk/ohb

NHS
Orkney

Dr Jim Wilson
Public Health Sciences
University of Edinburgh
4th Floor
MRC Human Genetics Unit
Western General Hospital
Crewe Road
Edinburgh
EH4 2XI

Date 27 February 2004

Tel: 01856 885421
Fax: 01856 885411
Enquiries to: Mrs Jean Aim
Email: jean.aim@orkney-hb.scot.nhs.uk

Dear Dr Wilson

Genetic Analysis of Complex Disease-Related Traits in Orkney:

The Research Ethics Committee initially reviewed the above application at the meeting held on 12 December 2003 and again on 13 January 2004 when further information was provided on a number of queries.

Ethical Opinion:

The members of the Committee present gave a favourable ethical opinion to the above research on the basis described in the application form, protocol and supporting documentation.

The favourable opinion applies to the NHS Orkney.

Principal Investigator – Dr James Flett Wilson

Management approval has also been given to this proposal.

Yours sincerely

Keith Farrer
Chairman – Orkney Local Research Ethics Committee

Runs of Homozygosity in European Populations

Ruth McQuillan,[1] Anne-Louise Leutenegger,[2] Rehab Abdel-Rahman,[1,7] Christopher S. Franklin,[1] Marijana Pericic,[3] Lovorka Barac-Lauc,[3] Nina Smolej-Narancic,[3] Branka Janicijevic,[3] Ozren Polasek,[1,4] Albert Tenesa,[5] Andrew K. MacLeod,[6] Susan M. Farrington,[5] Pavao Rudan,[3] Caroline Hayward,[7] Veronique Vitart,[7] Igor Rudan,[1,8,9] Sarah H. Wild,[1] Malcolm G. Dunlop,[5] Alan F. Wright,[7] Harry Campbell,[1] and James F. Wilson[1,*]

Estimating individual genome-wide autozygosity is important both in the identification of recessive disease variants via homozygosity mapping and in the investigation of the effects of genome-wide homozygosity on traits of biomedical importance. Approaches have tended to involve either single-point estimates or rather complex multipoint methods of inferring individual autozygosity, all on the basis of limited marker data. Now, with the availability of high-density genome scans, a multipoint, observational method of estimating individual autozygosity is possible. Using data from a 300,000 SNP panel in 2618 individuals from two isolated and two more-cosmopolitan populations of European origin, we explore the potential of estimating individual autozygosity from data on runs of homozygosity (ROHs). Termed F_{roh}, this is defined as the proportion of the autosomal genome in runs of homozygosity above a specified length. Mean F_{roh} distinguishes clearly between subpopulations classified in terms of grandparental endogamy and population size. With the use of good pedigree data for one of the populations (Orkney), F_{roh} was found to correlate strongly with the inbreeding coefficient estimated from pedigrees ($r = 0.86$). Using pedigrees to identify individuals with no shared maternal and paternal ancestors in five, and probably at least ten, generations, we show that ROHs measuring up to 4 Mb are common in demonstrably outbred individuals. Given the stochastic variation in ROH number, length, and location and the fact that ROHs are important whether ancient or recent in origin, approaches such as this will provide a more useful description of genomic autozygosity than has hitherto been possible.

Introduction

In plant and animal genetics, the detrimental effects of parental relatedness on fitness have long been recognized.[1] The mechanism of these effects is thought to be increased levels of homozygosity for deleterious recessive alleles, although overdominance might also play a role.[2]

In human populations in which consanguinity is customary or population size and isolation result in elevated levels of background parental relatedness, evidence has been reported of several effects, including an increased risk of monogenic disorders,[3–5]an increased risk of complex diseases involving recessive variants with intermediate or large effect sizes,[6–9] and genome-wide effects on disease traits such as blood pressure[10–17]and LDL cholesterol.[15] These are consistent with a causal role for many recessive variants with individually small effects scattered throughout the genome.

Central to any investigation of the effects of parental relatedness on the health of offspring is the need for a reliable and accurate method of quantifying this phenomenon at an individual level. The first method proposed was the inbreeding coefficient, F, defined as the probability of inheriting two identical-by-descent (IBD) alleles at an autosomal locus or, equivalently, the average proportion of the autosomal genome that is inherited IBD.[18] This is estimated with Wright's path method,[19] which calculates an individual's probability of inheriting two IBD alleles, given a specified pedigree and given that an allele present in a parent is transmitted to a specified offspring with a probability of 0.5. Before the availability of marker data from high-density genome scans, researchers had no option but to use this approach, despite the fact that, even where pedigrees are known and accurate, it has two major disadvantages.[20]

First, meiosis is a highly random process. Whereas on average, half of the DNA making up a gamete is maternally derived and half is paternally derived, there is a high degree of stochastic variance about this average.[21,22] As a consequence, grandchildren vary in the proportion of DNA they inherit from each of their four grandparents, and although the mean F coefficient of the offspring of first cousins is 0.0625, the standard deviation is 0.0243.[20] This variance increases with each meiosis (i.e., each degree of cousinship), so it is perfectly possible for the offspring of third cousins to be more autozygous (homozygous by descent) than the offspring of second cousins. Because the F coefficient (denoted here as F_{ped} to distinguish it from genomic estimates of autozygosity) is derived on the basis of this expectation, it is, therefore, only a very approximate estimate of individual genome-wide autozygosity.

[1]Public Health Sciences, University of Edinburgh Medical School, Edinburgh EH8 9AG, UK; [2]Unité de Recherche en Génétique Epidémiologique et Structure des Populations Humaines, INSERM U535, BP 1000, 94817 Villejuif, France; [3]Institute for Anthropological Research, 10000 Zagreb, Croatia; [4]Andrija Stampar School of Public Health, Faculty of Medicine, University of Zagreb, 10000 Zagreb, Croatia; [5]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Edinburgh EH4 2XU, UK; [6]Medical Genetics Section, University of Edinburgh, Molecular Medicine Centre, Edinburgh EH4 2XU, UK; [7]MRC Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, UK; [8]Croatian Centre for Global Health, Faculty of Medicine, University of Split, 21000 Split, Croatia; [9]Institute for Clinical Medical Research, University Hospital "Sestre Milosrdnice," HR-10000 Zagreb, Croatia
*Correspondence: jim.wilson@hgu.mrc.ac.uk
DOI 10.1016/j.ajhg.2008.08.007.

Second, F_{ped} estimates the proportion of an individual's genome that is IBD, relative to that of a poorly characterized founder generation. This generation is usually fairly recent, and, moreover, the founders are presumed to be unrelated, when in fact, members of historical populations were often related several times over through multiple lines of descent. As a result, this approach fails to capture the effects of distant parental relationships and, therefore, underestimates autozygosity, particularly in small, isolated populations or in populations with a long tradition of consanguineous marriage.[23,24]

With the increasing availability of high-density genome-scan data, interest has grown in exploring whether a more reliable and accurate estimate of autozygosity might be derived on the basis of genomic marker data. Much of the impetus for this comes from those searching for specific disease genes via homozygosity mapping, rather than from a general interest in the health effects of parental relatedness. Since the 1980s, many autosomal-recessive genes underlying monogenic human diseases have been identified with homozygosity mapping, which exploits the fact that regions flanking the disease gene will be identical by descent (IBD) in people with the disease whose parents are related to each other.[25] Botstein and Risch identified nearly 200 studies, published between 1995 and 2003, that used homozygosity mapping in consanguineous families to identify rare recessive disease genes.[26] Homozygosity mapping requires an estimate of the proportion of the genome that is autozygous for each affected individual, on the basis of which a LOD score for linkage to a specified locus is computed. Accurate estimation of autozygosity is crucial: underestimation results in an inflated LOD score and, thus, false evidence for linkage,[27,28] and overestimation results in false negatives.

Quantification of individual autozygosity is also of interest to those investigating recessive effects in complex-disease genetics. Several studies in consanguineous or small, isolated populations with above average levels of parental relatedness have found evidence for a genome-wide effect of homozygosity on coronary heart disease,[29–31] cancer,[29,32–34] blood pressure,[10–17] and LDL cholesterol.[15] These findings are consistent with studies suggesting that the variants associated with increased risk of common complex disease are more likely to be rare than to be common in the population;[35,36] are more likely to be distributed abundantly rather than sparsely across the genome,[37] and are more likely to be recessive than to be dominant.[38] Further empirical development of this idea has, however, been hampered by the inadequacy of available measures of autozygosity.

Here, we describe a multipoint, observational approach to estimating autozygosity from genomic data that exploits the fact that autozygous genotypes are not evenly distributed throughout the genome but are distributed in runs or tracts (Figure 1). This idea was first suggested by Broman and Weber, who proposed identifying autozygous segments from runs of consecutive homozygous markers.[39] Can runs of homozygosity (ROHs), observable from high-density genome-scan data, be used for a reliable and accurate estimate of autozygosity at both the individual level and the population level? How do individuals with different ancestry, characterized in terms of population size, endogamy, and parental relatedness, differ in terms of ROHs? At a population level, do ROHs reflect differences in population isolation?

This paper has three objectives. First, it uses various measures derived from ROHs to compare four European populations: two isolated island populations and two more-cosmopolitan populations. The key study population is the Scottish isolate of Orkney, a remote archipelago off the north coast of Scotland. Three additional populations are used for comparison: a representative Scottish comparison population,[40] an isolate population from a Dalmatian island in Croatia,[15] and the HapMap CEU (northwest-European-derived population from Utah, USA) founders from the Centre d'Étude du Polymorphisme Humain (CEPH).[41] Second, with the use of high-quality pedigree information available for the Orkney population, correlations are reported between F_{ped} and a genome-wide autozygosity measure derived from ROHs (F_{roh}). Finally, this study assesses the utility of F_{roh} as a measure of autozygosity.

Subjects and Methods

Study Populations

The Orkney Complex Disease Study (ORCADES) is an ongoing, family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the population isolate of the Orkney Isles in northern Scotland. The North Isles of Orkney, the focus of this study, consist of a subgroup of ten inhabited islands with census populations varying from ~30 to ~600 people on each island. Although transport links have steadily improved between the North Isles and the rest of Orkney, the geographical position of these islands, coupled with weather and sea conditions, means that even today they are isolated and that they would have been considerably more so in the past.

Although consanguinity is not the cultural norm in Orkney—indeed, there is evidence of consanguinity avoidance during the twentieth century[42]—two key factors make the North Isles population ideal for this type of study. First, the North Isles have experienced a period of severe population decline over the last 150 years, fueled by high emigration and low fertility. The population fell from an estimated peak of 7700 in the 1860s to 2217 by 2001. Second, endogamous marriage was widespread during the nineteenth century and into the twentieth centuries.[43] Therefore, despite consanguinity avoidance, the combined effects of steep population decline and endogamy have led to inflated levels of parental relatedness in the current population.

ORCADES received ethical approval from the appropriate research ethics committees in 2004. Data collection was carried out in Orkney between 2005 and 2007. Informed consent and blood samples were provided by 1019 Orcadian volunteers who had at least one grandparent from the North Isles of Orkney.

A Scottish comparison population was derived from the controls of the Scottish Colon Cancer Study (SOCCS).[40] This consists of 984 subjects, not known to have colon cancer, matched by

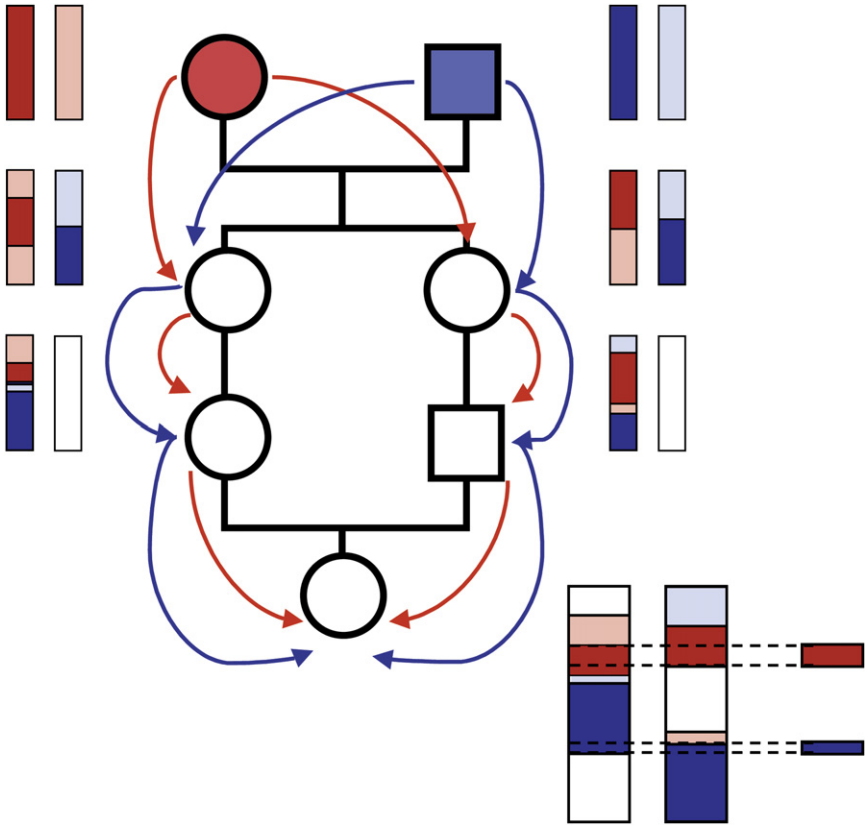


Figure 1. Pedigree of the Offspring of First Cousins

An example chromosome is illustrated. The female common ancestor is red. The chromosome inherited from one of her parents is colored red, and the chromosome inherited from her other parent is colored pink. The male common ancestor is blue. The chromosome inherited from one of his parents is colored dark blue, and the chromosome inherited from his other parent is colored light blue. The second generation are sisters. They share around 50% of their chromosomes IBD. The segments colored red and pink are segments inherited from their mother, and the segments colored dark and light blue are segments inherited from their father. The third generation are first cousins. In each case, the second (white) chromosome derives from their fathers (not shown), the red and pink segments are inherited from their maternal grandmother, and the dark and light blue segments are inherited from their maternal grandfather. The offspring of these first cousins has segments inherited from both founders on both copies of the chromosome. Where the same segments have been passed down both sides of the pedigree, the offspring of first cousins has extended identical-by-descent tracts or runs of homozygosity.

residential postal area and age to a series of incident cases of colorectal cancer. Subjects were resident throughout Scotland, with dates of birth ranging from 1921 to 1983.

The Dalmatian sample consists of 849 Croatian individuals, aged 18–93, sampled from the population of one island.[15] Both the SOCCS and the Croatian projects were approved by the relevant ethics committees.

The CEU sample consists of 60 unrelated individuals from Utah, USA, of northwest-European ancestry, collected by the CEPH in 1980.[41]

Genotyping

Genotyping procedures for the Scottish,[40] Dalmatian,[44] and CEU[45] samples are described elsewhere. All were genotyped on the Illumina Infinium HumanHap300v2 platform (Illumina, San Diego, CA, USA). After extraction of genomic DNA from whole blood with the use of Nucleon kits (Tepnel, Manchester, UK), 758 Orcadian samples were genotyped, according to the manufacturer's instructions, on the Illumina Infinium HumanHap300v2 platform. Analysis of the raw data was done via BeadStudio software, with the recommended parameters for the Infinium assay, with the use of the genotype-cluster files provided by Illumina.

Individuals with less than 95% call rate were removed, as were SNPs with more than 10% missing genotypes. SNPs failing Hardy-Weinberg equilibrium at a threshold of 0.0001 were removed. IBD sharing between all first- and second-degree relative pairs was assessed with the *Genome* program in PLINK,[46] and individuals falling outside expected ranges were removed from the study. Sex checking was performed with PLINK, and individuals with discordant pedigree and genomic data were removed. On completion of data-cleaning and quality-control procedures, 725 individuals and 316,364 autosomal SNPs remained. The male-to-female ratio of study participants is 0.86. The mean year of birth is 1952, varying from 1909 to 1988.

A consensus SNP panel was then created, with use of only those markers that satisfied these quality control criteria in all four populations, leaving a final sample of 289,738 autosomal SNPs and 2618 individuals (60 from CEU, 725 from Orkney, 849 from the Dalmatian island, and 984 from Scotland).

F_{ped} Estimates

The pedigrees of all individuals in the ORCADES sample were traced back for as many generations as possible in all ancestral lineages, with the use of official birth, marriage, death, and census records held by the General Register Office for Scotland in Edinburgh. F_{ped} was calculated for each individual via Wright's path method.[19]

Limited pedigree information is available for the Dalmatian-isolate data set, and this is too incomplete for an estimate of F_{ped}. It was, however, possible to analyze these data with the use of grandparental-endogamy levels.

No pedigree information is available for the Scotland data set; however, we analyzed data according to the rurality of subjects' residential address[47] in order to determine whether there is any evidence for an association between remote rurality and autozygosity in Scotland.

Runs of Homozygosity

ROHs were identified via the Runs of Homozygosity program implemented in PLINK version 1.0.[46] This slides a moving window of

5000 kb (minimum 50 SNPs) across the genome to detect long contiguous runs of homozygous genotypes. An occasional genotyping error or missing genotype occurring in an otherwise-unbroken homozygous segment could result in the underestimation of ROHs. To address this, the program allows one heterozygous and five missing calls per window.

A threshold was set for the minimum length (kb) needed for a tract to qualify as homozygous. Because strong linkage disequilibrium (LD), typically extending up to about 100 kb, is common throughout the genome,[48–51] short tracts of homozygosity are very prevalent. For exclusion of these short and very common ROHs that occur in all individuals in all populations, the minimum length for an ROH was set at 500 kb. All empirical studies have identified a few very long stretches of LD, measuring up to several hundred kb in length,[49] which could result in the occurrence of longer ROHs in outbred individuals. Such ROHs will not be excluded by this methodology; however, the purpose here is not to identify only those ROHs that result from parental relatedness but to identify all ROHs and then relate these to pedigree and population data for an assessment of the extent to which these result from parental relatedness and population isolation.

We set a threshold for the minimum number of SNPs constituting a ROH in order to ensure that these are true ROHs—i.e., that between the first SNP and the last SNP the entire unobserved stretch of the chromosome is homozygous. With, for example, only three consecutive homozygous genotypes, there would be a very high probability that these three could be homozygous by chance alone and that the intervening, unobserved chromosomal stretches could be heterozygous. We have deliberately not taken LD into account here. By using a minimum-length cutoff of 500 kb, most shorter ROHs resulting from LD will be eliminated; however, some longer stretches will remain. This is intentional: we are interested in identifying and quantifying these common ROHs, whatever their origin. We used allele frequencies for a random sample of chromosomal segments across the entire autosomes to estimate the mean probability of finding 10, 25, and 50 consecutive homozygous SNPs by chance alone in each population. On this basis, the minimum number of contiguous homozygous SNPs constituting a ROH was set at 25 ($p < 0.0001$ in each of the four populations). Two additional parameters were added for ensuring that estimates of F were not artificially inflated by apparently homozygous tracts in sparsely covered genomic regions: tracts with a mean tract density > 50 kb/SNP were excluded, and the maximum gap between two consecutive homozygous SNPs was set at 100 kb.

For exclusion of the possibility that apparent ROHs are in fact regions of hemizygous deletion, an analysis of deletions was carried out in the Orkney data set. An Objective Bayes' Hidden Markov model, as employed in QuantiSNP v. 1.0, was used for identification of heterozygous deletions with a sliding window of 2 Mb over the genome and 25 iterations. All of the samples were corrected for genomic GC content prior to copy-number inference as a means of ensuring that the variation of the observed $\log_2$ R ratio is not attributed to the region-specific GC content.[52] We included in the downstream analysis all heterozygous deletions with an estimated Bayes' factor ≥ 10 to ensure a low false-negative rate, as reported in Colella et al., 2007.[53] A custom Perl script was developed for comparison of the identified heterozygous deletions and ROHs.

All deletions overlapping with ROHs were identified. When deletions covered the entire length of the ROH or when less than 0.5 Mb of the tract remained after the deletion was taken account of, the ROH was removed from the analysis. Because the Dalmatian, CEU, and Scotland data sets were uncorrected for deletions, uncorrected Orkney data are shown when there are population comparisons. Analyses using only the Orkney data set use data corrected for deletions.

F_{roh} Estimates

A genomic measure of individual autozygosity (F_{roh}) was derived, defined as the proportion of the autosomal genome in runs of homozygosity above a specified length threshold:

$$F_{roh} = \sum L_{roh}/L_{auto}$$

in which $\sum L_{roh}$ is the total length of all of an individual's ROHs above a specified minimum length and L_{auto} is the length of the autosomal genome covered by SNPs, excluding the centromeres. The centromeres are excluded because they are long genomic stretches devoid of SNPs and their inclusion might inflate estimates of autozygosity if both flanking SNPs are homozygous. The length of the autosomal genome covered by our consensus panel of SNPs is 2,673,768 kb. We show individual and population mean values of F_{roh} for a range of different ROH-length thresholds.

Statistical Analysis

For statistical analyses, the Orkney population was split into endogamous Orcadians, defined as those with at least three grandparents born in Orkney, on the same island, typically ~10 km^2 in size and with a population of 50–500 ($n = 390$); mixed Orcadians, defined as those with at least three grandparents born in Orkney but on different islands in the archipelago—i.e., from an area over 500 km^2 with a population of ~20,000 ($n = 286$); and half Orcadians, defined as those with one pair of Orcadian-born and one pair of Scottish-mainland-born grandparents ($n = 49$). Although pedigree information is not available for an assessment of whether the parents of half-Orcadian subjects are related beyond five generations in the past, it is reasonable to assume that they are likely to be unrelated for at least 10–12 generations. It is known that there was major Scottish immigration to Orkney in the 15th and 16th centuries, before10–12 generations ago. Although Scottish immigration has certainly occurred sporadically since then, rates have been low. An analysis of the area of origin of the Scottish parents of our half-Orcadian subjects shows that they came from all over Scotland: we found no evidence for strong Orcadian connections with any specific Scottish settlement, which might increase the chances of parental relatedness in this group. Furthermore, the surnames of the ancestors of the Orcadian parents of this group were markedly different from those of the ancestors of the non-Orcadian Scottish parents.

The Dalmatian population was split into endogamous Dalmatians, defined as those with all four grandparents born in the same village—i.e., from a 1 km^2 area, with a population of $<$ 2000 ($n = 431$); mixed Dalmatian, defined as those with all four grandparents born on the same island but not in the same village—i.e., from a 90 km^2 area with a population of 3600 ($n = 221$); and Croatian, defined as residents of the island with grandparents born elsewhere in Croatia ($n = 197$). The CEU and Scottish populations were not subdivided.

All calculations were performed with SPSS and Excel software. The proportions of each subpopulation with ROHs measuring less than 1, 1.5, and 2 Mb were calculated. All subjects in all subpopulations had ROHs shorter than 1.5 Mb. Subpopulations start to become differentiated from each other for ROHs > 1.5 Mb, with the effects of endogamy on ROHs starting to emerge above this

threshold. Unless otherwise specified, all analyses exploring the effects of endogamy and parental relatedness on ROHs therefore define a ROH as measuring $\geq$ 1.5 Mb.

Subpopulation means were calculated for the total length of ROHs per individual. The number of ROHs was plotted against the total length of ROHs, per individual, for each subpopulation.

The correlation between F_{ped} and F_{roh} was calculated with the use of a subset of 249 individuals, from the Orkney sample, who satisfied the condition of having at least two grandparents on the same side of the family born in Orkney and no grandparents born outside of Scotland and who were either the offspring of consanguineous parents (parents related as 2nd cousins or closer) or those for whom it was possible to establish pedigrees for at least six generations in all Orcadian ancestral lineages or five generations in non-Orcadian ancestral lineages.

Correlations were also calculated between F_{roh}, F_{ped}, and two other measures: multilocus heterozgyosity (MLH), which is defined as the proportion of markers that are heterozygous,[54] and the measure of autozygosity implemented in PLINK, termed here F_{plink}, which estimates autozygosity from genotype frequencies, giving more weight to rare alleles.[46]

Prevalence and Genomic Location of ROHs in Different Subpopulations

Next, we explored the hypothesis that ROHs in outbred individuals tend to cluster in the same genomic locations, whereas those present in the offspring of related parents tend to be more randomly distributed across the autosomes. We compared the location of ROHs in three groups: the half-Orcadian group, consisting of all half Orcadians with at least one ROH measuring $\geq$ 1.5 Mb (n = 46); an offspring-of-cousins group, which was constructed by consideration of all individuals from the Orkney sample with parents related as 3rd cousins or closer and the selection of those 20 with the greatest total length of ROHs; and a control population derived from our cross-sectional sample from Scotland. Because some individuals in the Scottish sample have long ROHs that could be indicative of parental relatedness, we restricted the control sample to those with no more than eight ROHs, totaling no more than 17 Mb: the maximum values in the half-Orcadian group, the members of which are known to be the offspring of unrelated parents. There were 943 individuals in the control group. ROHs measuring at least 1.5 Mb in all three groups were compared. Control-group ROHs overlapping by at least 0.5 Mb with ROHs in either Orcadian group were counted. The number of control overlaps per ROH (and per Mb of ROH) in the half-Orcadian group was compared with that in the offspring-of-cousins group.

We then investigated whether ROHs in half Orcadians occurred in regions of lower-than-average recombination. Based on sex-averaged mean recombination rates per Mb, derived from the deCODE genetic map, we used the UCSC Genome Browser (March 2006)[55] to calculate the mean recombination rate of all complete Mb of ROH in our half-Orcadian sample.

Results

Copy-Number Variation

We detected 224 deletions that overlapped with ROHs (median length of deletion 995 kb). Overlapping deletions were detected in 57 individuals (7.6% of sample). After removal of these overlaps from the sample and removal of the entire affected ROH if less than 0.5 Mb remained, ROH statistics were recalculated. There was no significant difference between results before and after correction for deletion for the mean total length of ROHs (correcting for deletions reduced this by less than 0.3% in the sample as a whole) or the mean number of ROHs (reduced by 0.02%). Furthermore, no significant differences were found when data were analyzed by subpopulation and when different length parameters were used for defining ROHs. This provides strong evidence that the ROHs identified are true homozygous tracts and not hemizgyous deletions.

Urban versus Rural Analysis of Scottish Sample

No difference was found in the mean total length of ROHs between those living in rural areas and those in urban areas of Scotland, regardless of whether the analysis used a dichotomous classification or a more-detailed, eight-category classification, from large urban to remote rural (data not shown). Data were also analyzed for a subset (n = 426) of the sample with information on grandparental country of birth. On average, those with four Scottish-born grandparents (n = 254) had a slightly greater sum of ROHs than did those with at least one grandparent born outside of Scotland, but differences were not significant (data not shown). The Scottish sample was, therefore, not split into subpopulations for further analyses.

Effect of Stochastic Variation on Individual Autozygosity

On average, the difference in the total length of ROHs between full sibling pairs was 10.3 Mb. However, the distribution is skewed, with half of all individuals having less than 5 Mb difference yet some 7% differing by more than 30 Mb. The greatest difference between sibling pairs was 91 Mb, or 3.4% of the autosomes (paternity was confirmed from patterns of genomic sharing in all cases).

Effects of Population Isolation and Endogamy on Length and Number of ROHs

The proportions of subpopulations with ROHs of a given length are shown in Figure 2. All individuals in all populations have ROHs measuring less than 1.5 Mb. If we consider the populations as a whole, on average, a significantly greater proportion of the autosomes of Orcadians are in ROHs measuring 0.5–1.5 Mb (77.7 Mb) than is the case for either the Dalmatian (73.2 Mb), the Scottish (75.8 Mb), or the CEU (74.1 Mb) populations. There are no significant differences between groups within populations, however, which suggests that this reflects population differences in genetic diversity or LD of ancient origin rather than effects of more recent endogamy or population isolation.

For ROHs above 1.5 Mb, three distinct groupings, which are clearly related to endogamy and isolation, emerge: a greater proportion of the endogamous Dalmatian and Orcadian samples than of the other samples have long ROHs (28% have ROHs > 10 Mb); only a small proportion of the CEU, Scottish, and half-Orcadian samples have long

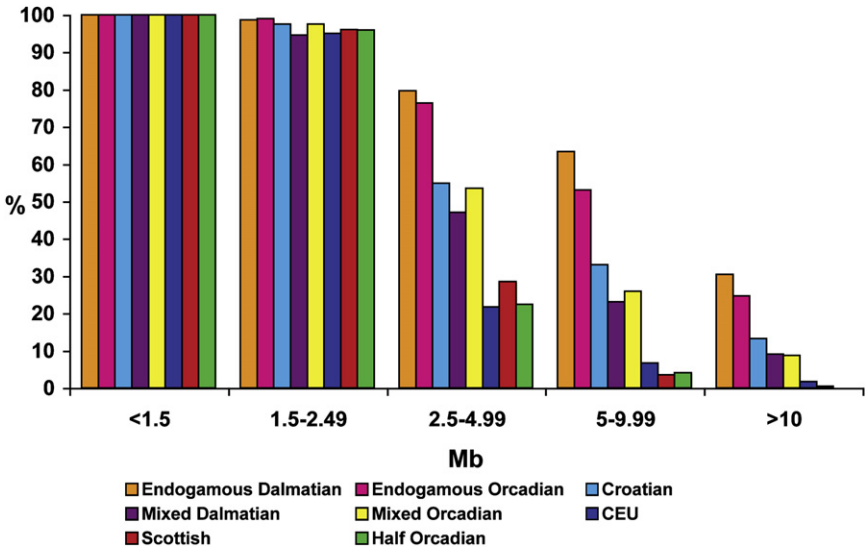


Figure 2. Proportion of Subpopulations with One or More ROHs of a Given Length
The proportion of individuals with one or more ROHs of up to 0.5–1.49, 1.5–2.49, 2.5–4.99, and 5–9.99 Mb in length, or over 10 Mb in length, is plotted for each of the eight population groups defined in the Statistical Analysis section of Subjects and Methods.

ROHs (0.5% > 10 Mb), and the proportion of Croatian and mixed Dalmatian and Orcadian samples with long ROHs falls in between (10% > 10 Mb).

Forty-nine individuals had no ROHs longer than 1.5 Mb. This number included at least one individual from each subpopulation, although they were predominantly half-Orcadian, Scottish, and CEU samples. The shortest sum of ROHs across all of the samples was found in a Scottish individual, who had ROHs longer than 0.5 Mb covering only 1.5% of the autosomes (39 Mb). This compares with a mean of 3.5% across all of the populations (93 Mb).

The number of ROHs longer than 1.5 Mb per individual, plotted against the total length of those ROHs, is shown for each group in Figure 3. The half-Orcadian group is used as a reference, because we know that these individuals are the offspring of unrelated parents. Reference lines are shown on all graphs for the maximum number of ROHs, the maximum total length of ROHs, and the line of best fit for the half-Orcadian group. Compared with the half-Orcadian group, all other groups have a greater variance in the number and sum of ROHs and contain individuals with more and longer ROHs. Again, the same three groupings are apparent. Data points for the half-Orcadian, Scottish, and CEU samples are generally narrowly distributed along both axes, indicating that these individuals have few, relatively short ROHs. The two endogamous samples are much more widely spread along both axes, reflecting the presence of many, much longer ROHs. The Croatian, mixed Orcadian, and mixed Dalmatian groups are intermediate, reflecting the fact that these less carefully specified groups are probably made up of individuals with a mixture of ancestries, from the outbred to the very endogamous. The percentage of each group with more and longer ROHs than the maximum for the half Orcadians was calculated. Again, the Scottish (5%) and CEU (8%) groups differed least and the endogamous Dalmatians (64%) and Orcadians (54%) differed most from the half Orcadians. The Croatians (33%), mixed Dalmatians (26%), and mixed Orcadians (23%) were intermediate.

The effect of different degrees of parental relatedness on the sum and number of ROHs is shown in Figure 4 for the 249 individuals in the Orkney sample with good pedigree information. Although a trend for increasing number and total length of ROHs is evident from the half-Orcadian through the mixed to the endogamous and offspring-or-cousins subgroups, there is considerable overlap between groups.

Comparison of F_{ped} and F_{roh}

A subset of 249 Orcadian individuals with complete and reliable pedigree data were used to compare F_{ped} and F_{roh}. The mean (standard error) F_{ped} of the sample is 0.0038 (0.0005), approximately equivalent to a parental relationship of third cousins. Mean F_{ped} values for Orcadian subpopulations are shown in Table 1. These vary from 0.02, for the offspring of 1st or 2nd cousins, to 0.0002 (equivalent to a parental relationship of 5th cousins) in the mixed Orcadian group. Mean F_{ped} values are compared with mean F_{roh} values for a range of minimum-length thresholds. The mean value of $F_{roh\ 5}$ (i.e., with a minimum-length threshold of 5 Mb) is closest to that of F_{ped}, whereas $F_{roh\ 0.5}$ (i.e., with a minimum-length threshold of 0.5 Mb) is an order of magnitude higher. This suggests that a shared maternal and paternal ancestor in the preceding six generations results predominantly in ROHs longer than 5 Mb. It is clear from the half-Orcadian group, whose parents do not share a common ancestor for at least six generations and probably at least 10–12 generations, that ROHs measuring less than 3 or 4 Mb are not uncommon in the absence of parental relatedness. On average, these individuals have over 3% (84 Mb) of their autosomes in ROHs over 0.5 Mb long and 0.2% (almost 6 Mb) in ROHs longer than 1.5 Mb.

Correlation between F_{roh}, F_{ped}, F_{plink}, and MLH

We used the total sample to examine correlations between different genetic estimates of autozygosity or homozygosity. Because MLH is in fact a measure of heterozygosity, we have used 1 − MLH in our calculations. Allele frequencies for F_{plink} were estimated by naive counting in all individuals, as implemented in PLINK. F_{plink} and 1 − MLH are

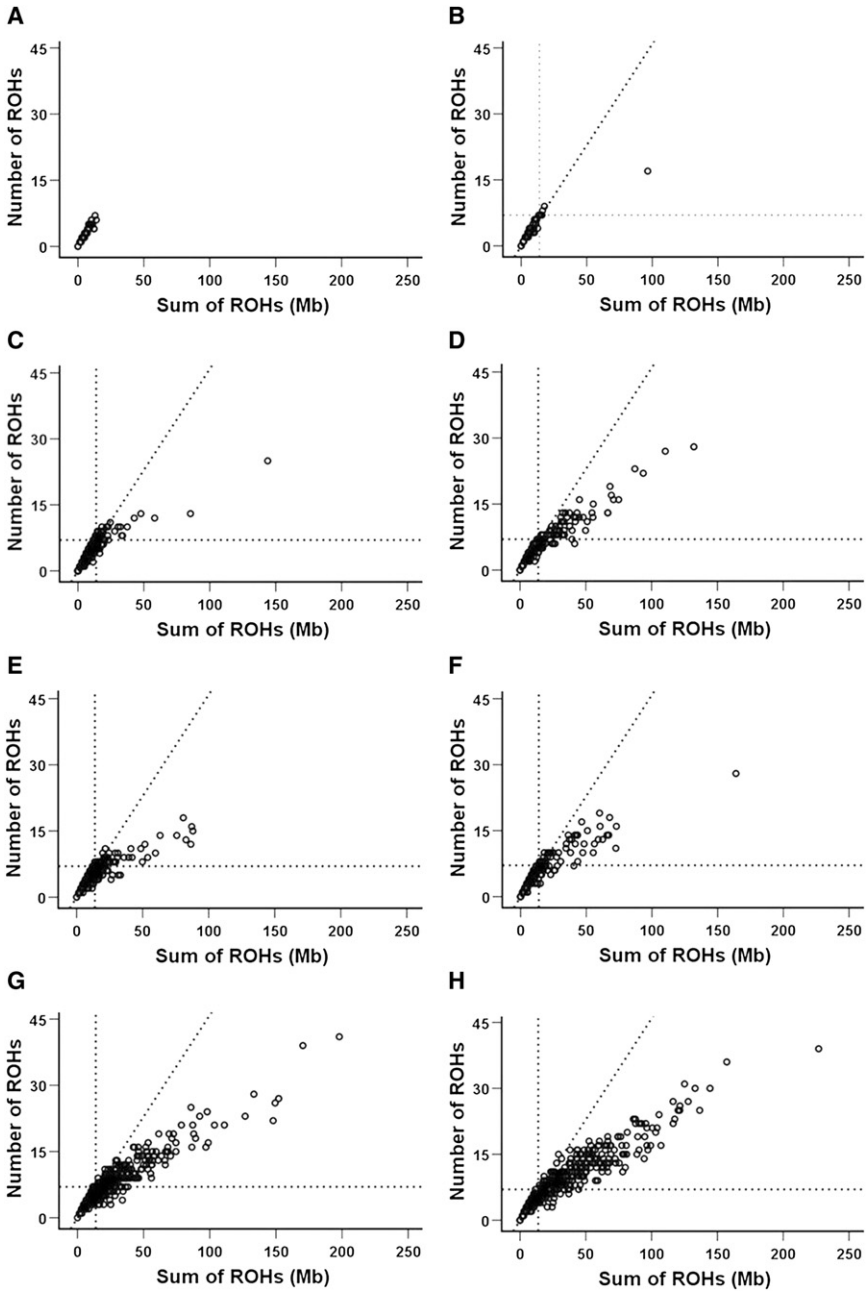


Figure 3. Number of ROHs Compared to Total Length of ROHs
(A) Half Orcadian, (B) CEU, (C) Scottish, (D) Croatian, (E) Mixed Orcadian, (F) Mixed Dalmatian, (G) Endogamous Orcadian, and (H) Endogamous Dalmatian.

highly correlated (r = 0.94). $F_{roh\ 1.5}$ is more highly correlated with 1 − MLH (r = 0.80) than with F_{plink} (r = 0.74).

We used a subset of the Orcadian sample (n = 249) to estimate correlations with F_{ped}. $F_{roh\ 1.5}$ was most highly correlated with F_{ped} (r = 0.86; 95% confidence interval 0.83–0.89). Correlations between F_{ped} and $F_{roh\ 1.5}$ were significantly higher than both the correlation between F_{plink} and F_{ped} (r = 0.77; 0.72–0.82) and that between 1 − MLH and F_{ped} (r = 0.76; 0.71–0.82). $F_{roh\ 1.5}$ was slightly, but not significantly, more strongly correlated with F_{ped} than was either $F_{roh\ 0.5}$ or $F_{roh\ 5}$.

Correlations between F_{ped} and $F_{roh\ 0.5}$, $F_{roh1.5}$, and $F_{roh\ 5}$ are shown in Figure 5. For each value of F_{ped} there is a range of values for F_{roh}, reflecting stochastic variation in ancestral recombination, the existence of multiple distant parental relationships undetectable with the use of pedigrees, and possible pedigree misspecifications. The closer the parental relationship, the greater the variance in the autozygosity of offspring. This is clear from the wide distribution of F_{roh} values in the endogamous group compared to the mixed Orcadian group. Although as we have shown, ROHs shorter than around 1.5 Mb do not appear to reflect differences in recent ancestral endogamy, data from the half-Orcadian sample illustrate that the prevalence of these shorter ROHs clearly varies between individuals. Use of a minimum-ROH-length threshold of 5 Mb might better reflect the effects of parental relatedness on autozygosity; however, it also obscures a great deal of individual genetic variation of more ancient origin. This is illustrated by the regression lines on each panel: the y intercept gives the value of F_{roh} when $F_{ped} = 0$. This is a measure of the proportion of the autosomes in ROHs not captured by F_{ped}. Thus, 0.034 of the autosomes are in ROHs longer than 0.5 Mb but are not captured by F_{ped}. The equivalent figures are 0.0053 for ROHs longer than 1.5 Mb and 0.0014 for ROHs longer than 5 Mb. This clearly shows that F_{ped} fails to account for autozygosity of ancient origin.

Mean F_{roh} by Subpopulation

Mean F_{roh} and the mean total length of ROHs for each subpopulation are shown for a range of minimum ROH lengths in Figure 6. This figure again shows the effect on F_{roh}, in all populations, of changing the ROH-length cutoff point. The same three distinct groupings emerge for ROHs longer than 1.5 Mb, although when shorter ROHs are included, the picture is less clear. With 1.5 Mb used as the minimum length, endogamous Dalmatians have a mean F_{roh} of 0.013 (35 Mb), endogamous Orcadians 0.011 (28 Mb), Croatians 0.007 (18 Mb), mixed Dalmatians 0.006 (15 Mb), mixed Orcadians 0.005 (14 Mb), CEU 0.003

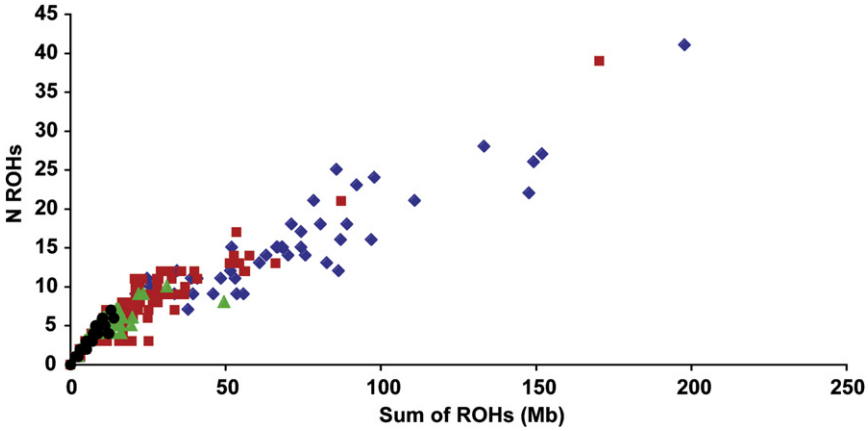


Figure 4. Effect of Endogamy on Sum and Number of ROHs
Offspring of 1st or 2nd cousins are shown in blue, endogamous Orcadians who are not the offspring of 1st or 2nd cousins are shown in red, mixed Orcadians are shown in green, and half Orcadians are shown in black.

(8 Mb), Scottish 0.003 (7 Mb), and half Orcadians 0.002 (6 Mb). With a 5 Mb threshold, the same relationship between groups is seen, but values for all groups are reduced (to 17 Mb in endogamous Dalmatians and 0.3 Mb in half Orcadians).

Comparison of ROHs in the Offspring of Unrelated Parents and the Offspring of Cousins

We next investigated whether ROHs found in half Orcadians are more common than those found in the offspring of related parents. We defined "common" as overlapping by at least 0.5 Mb with ROHs found in a subset of the Scottish sample. The number of ROHs measuring $\geq$ 1.5 Mb was 143 in the half-Orcadian sample, 3159 in the Scottish control sample, and 382 in the offspring-of-cousins sample. Results are summarized in Table 2. On average, each half-Orcadian ROH overlapped with more than twice as many controls as did ROHs in the offspring-of-cousins group. Only 12.6% of half-Orcadian ROHs, but almost a third of ROHs in the offspring-of-cousins group, did not overlap with any controls. We also looked at the mean number of overlaps per Mb of ROH in the two samples in order to correct for the fact that ROHs in the offspring-of-cousins group tend to be longer. There were more than three times as many control overlaps per Mb of ROH in the half-Orcadian group than there were in the offspring-of-cousins group. If we consider only those ROHs measuring $>$ 5 Mb in the offspring-of-cousins sample (i.e., those that are most likely to result from recent shared parental ancestry), the mean number of overlaps per Mb was only1.4 (SD 2.0).

Data on chromosome 1 for ten individuals in the half-Orcadian group (shown in blue) and seven individuals in the offspring-of-cousins group (shown in red) are illustrated by way of example in Figure 7. These are all of the individuals in the sample with ROHs on chromosome 1, except that data for only one individual per sibship is shown. This removed six individuals from the offspring-of-cousins group but none from the half-Orcadian group. The numbers shown below each colored segment are the numbers of ROHs in the control sample overlapping with the illustrated ROH. It is clear that although there is a tendency for ROHs from both groups to cluster in certain chromosomal regions, the longer ROHs in the offspring-of-cousins group are more randomly distributed along the chromosome.

Next, we identified all ROHs in the half-Orcadian group that overlapped by at least 0.5 Mb with common ROHs identified by Lencz.[56] In a sample of 322 non-Hispanic European Americans, Lencz identified 339 ROHs present in at least ten subjects. Of the 143 half-Orcadian ROHs, 57% overlapped with Lencz et al.'s list. Only 7% (ten ROHs) overlapped with neither Lencz et al.'s list nor our control group.

Finally, we investigated whether the ROHs in half Orcadians were found in areas of lower-than-average recombination. The mean recombination rate for the regions where half-Orcadian ROHs are located is 0.52 of the mean genome-wide recombination rate. For common ROHs (i.e., half-Orcadian ROHs that overlap with ROHs in the control group), this figure was 0.38 of the genome-wide mean.

Discussion

Our findings are consistent with a number of recent observational studies using high-density genome-scan data, which have suggested that ROHs longer than 1 Mb are more common in outbred individuals than previously thought.[39,56–60]

We have quantified this phenomenon by describing the number and length of ROHs in individuals who are known to have no common maternal and paternal ancestor in at least five generations (and probably 10–12 generations). Our analysis of copy-number variation in the Orkney sample is consistent with studies that have shown that observed ROHs are true homozygous tracts and not deletions or other chromosomal abnormalities.[39,45,57,60] Heterozygous deletions are not easily differentiated from ROHs, because the employed algorithm uses the B allele frequency as one of its input parameters to infer CNV status. Therefore, homozygosity at consecutive SNPs increases the posterior probability of being called a heterozygous deletion. In other words, this is a very robust estimation of the prevalence of ROHs in the Orkney sample, which to some extent overcorrects for heterozygous deletions. Other studies have suggested that ROHs cluster in regions of the genome where recombination rates are low,[57–60] and our data

Table 1. Mean Values of F_{ped} and F_{roh} for Orkney Subpopulations

Orkney Subpopulation	N	Mean (SE) F_{ped}	Equivalent Parental Cousin Relationship (Single Loop)	Mean (SE) $F_{roh\ 0.5}$	Mean (SE) $F_{roh\ 1.5}$	Mean (SE) $F_{roh\ 5}$
Offspring of 1st or 2nd cousins	42	0.0182 (0.0014)	2nd cousin	0.0569 (0.0024)	0.0271 (0.0022)	0.0169 (0.0017)
Endogamous Orcadian	114	0.0015 (0.0004)	3rd – 4th cousin	0.0379 (0.0008)	0.0087 (0.0007)	0.003 (0.0004)
Mixed Orcadian	44	0.0002 (0.0001)	5th cousin	0.033 (0.0006)	0.0046 (0.0005)	0.0012 (0.0004)
Half Orcadian	49	0	None	0.0315 (0.0004)	0.0021 (0.0002)	0.0001 (0.00007)
Total	249	0.0038 (0.0005)	3rd cousin	0.039 (0.0008)	0.0098 (0.0007)	0.0045 (0.0005)

support this. The picture of genome-wide homozygosity now emerging is that short stretches, measuring tens of kb and indicative of ancient LD patterns, are common, covering up to one third of the genome.[45] At the other end of the spectrum, very long ROHs, measuring tens of Mb, are the signature of parental relatedness. In between, ROHs might result from recent parental relatedness or might be autozygous segments of much older pedigree that have occurred because of the chance inheritance through both parents of extended haplotypes that are at a high frequency in the general population, possibly because they convey or conveyed some selective advantage.[56] The Phase II HapMap study estimates that ROHs measuring in excess of around 100 kb constitute 13%–14% of the genome in Europeans.[45] Lencz et al.[56] give a similar estimate. The findings of our study are not directly comparable, given that we have not examined ROHs shorter than 500 kb; however, we have shown (Figure 2) that ROHs measuring between 500 and 1500 kb were present in all individuals in all the subpopulations that we studied, totaling on average 75 Mb per individual (2%–3% of the autosomes). The fact that we found small but significant differences *among* our four populations in the mean sum of these short ROHs but no significant differences *within* populations (e.g., between endogamous Orcadians and half Orcadians) lends support to the view that population differences in the prevalence of ROHs shorter than around 1.5 Mb reflect LD patterns of ancient origin rather than the effects of more recent endogamy.

We have demonstrated clearly that data on ROHs measuring more than 1.5 Mb accurately reflect differences in population isolation, as measured by grandparental endogamy (Figures 2, 3, and 6). Furthermore, characterizing populations in terms of ROHs allows us to situate those with unknown degrees of isolation along a spectrum. For example, beyond knowing that the Scottish sample is broadly representative of the general Scottish population, we have no information on the precise birthplace of participants' grandparents. Data on ROHs would suggest that endogamy and consanguinity are uncommon, although not unheard of, in the recent ancestry of modern Scots. The 36 (4%) outliers in Scottish sample with ROHs suggestive of parental relatedness (total ROHs $\geq$ 5 Mb) were no more likely to live in rural or island locations than in urban locations. This is unsurprising: Scotland is a small, largely urbanized country with high population mobility. There are, however, small, remote island communities off the west and north coasts of Scotland that have been shown

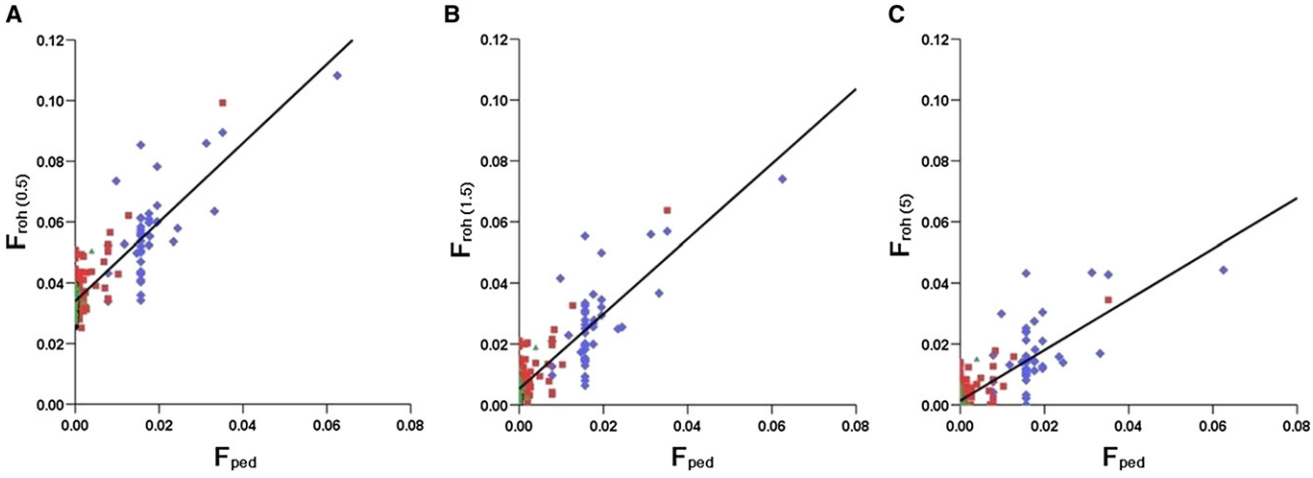


Figure 5. Correlation between F_{ped} and F_{roh} in Orkney Sample
Correlations, with regression lines, are shown for three different minimum-ROH-length thresholds. (A) shows the correlation between F_{ped} and $F_{roh\ 0.5}$, (B) shows the correlation between F_{ped} and $F_{roh\ 1.5}$, and (C) shows the correlation between F_{ped} and $F_{roh\ 5}$. For colors and details of subgroups, see Figure 4 legend. N = 249.

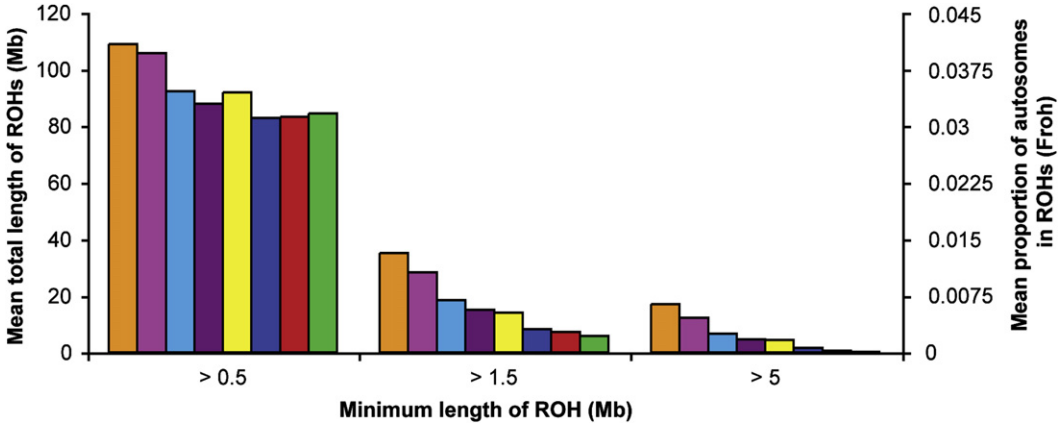


Figure 6. Mean Total Length of ROHs over a Range of Minimum Tract Lengths
The average total length of ROHs per individual, calculated from ROHs above 0.5, 1.5 and 5 Mb, is plotted for each of the eight population groups defined in the Statistical Analysis section of Subjects and Methods. For colors, see Figure 2 legend.

to have greater LD and lower haplotype diversity than mainland urban and rural Scottish populations,[61] consistent with lower effective population sizes, isolation, and genetic drift. Orkney is one such isolated community; however, as we show, even within such small populations, there is a great diversity of ancestry, from the tightly endogamous to the completely outbred. Our data show that having at least three grandparents from within a 2–3 mile radius (as is the case in the North Isles of Orkney and the Dalmatian villages) is associated with considerably more and longer ROHs than is merely coming from Orkney or a Dalmatian island. The distribution of ROHs in the CEU sample, which is widely used as a northwest-European reference population, does indeed appear to be very similar in this respect to that in the Scottish sample. Consistent with other studies,[45] we identify one outlier (NA12874), who is likely to be the offspring of consanguineous parents. The Dalmatian subsample of the offspring of Croatian settlers is more autozygous by various ROH-based measures than the mixed-Dalmatian and mixed-Orcadian subgroups, suggesting that these settlers came from fairly small, semi-isolated communities where endogamy was not uncommon.

Table 2. Overlaps between ROHs Found in Orcadians and Those Found in a Scottish Control Sample

	Half Orcadian	Offspring of Cousins
Number of individuals	46	20
Number of ROHs $\geq$ 1.5 Mb	143	382
Mean (SE) number of control overlaps per ROH	20.5 (22.5)	9.6 (16.0)
Maximum number of controls overlapping with a ROH	123	123
Percentage of ROHs overlapping with no controls	12.6	29
Mean (SE) number of control overlaps per Mb of ROH	10.9 (11.8)	3 (6.3)

We found that F_{roh} is strongly correlated with F_{ped}, significantly more so than the other two measures investigated. Perfect correlation is not expected, largely because of the deficiencies of F_{ped}. This is particularly the case in isolated populations, where multiple distant parental relationships, undetectable with only a few generations of pedigree information, inflate autozygosity, such that the offspring of distant cousins can be almost as autozygous as the offspring of first cousins.[24] The individual with the second highest F_{roh} in the Orkney sample, for example, is the offspring of a couple whose closest relationship is that of 3rd cousins but who are multiply related at least 24 different ways in the last eight generations alone. We illustrate the deficiencies of F_{ped} in Figure 5, in which the y intercept of the regression line is an indication of the autozygosity captured by F_{roh} but not by F_{ped}. Although it is unlikely that any approach could accurately identify the precise nature of distant parental cousin relationships for individuals with such complex pedigrees as those found in our Orkney sample, F_{roh} can accurately rule out the possibility that an individual is the offspring of first cousins: during preliminary data analysis, before all pedigree relationships had been verified by checking of inferred IBD sharing among first-degree relatives, a sibling pair, putatively the offspring of first cousins, was identified as having F_{roh} values significantly lower than predicted from pedigree. Upon checking of inferred IBD sharing among pairs of their genotyped relatives, an ancestral false paternity was identified that explained this anomaly.

A key objective of this research was to explore whether ROHs could be used for derivation of a measure of individual autozygosity. Before the advent of dense genome scans, the approach to estimating autozygosity from genetic-marker data was invariably inferential. We propose a very different, observational approach. Termed F_{roh}, this is defined as the proportion of the autosomal genome in ROHs above a specified length threshold. Our purpose here is not to develop a fully fledged statistical methodology tested against the alternatives—further work is needed

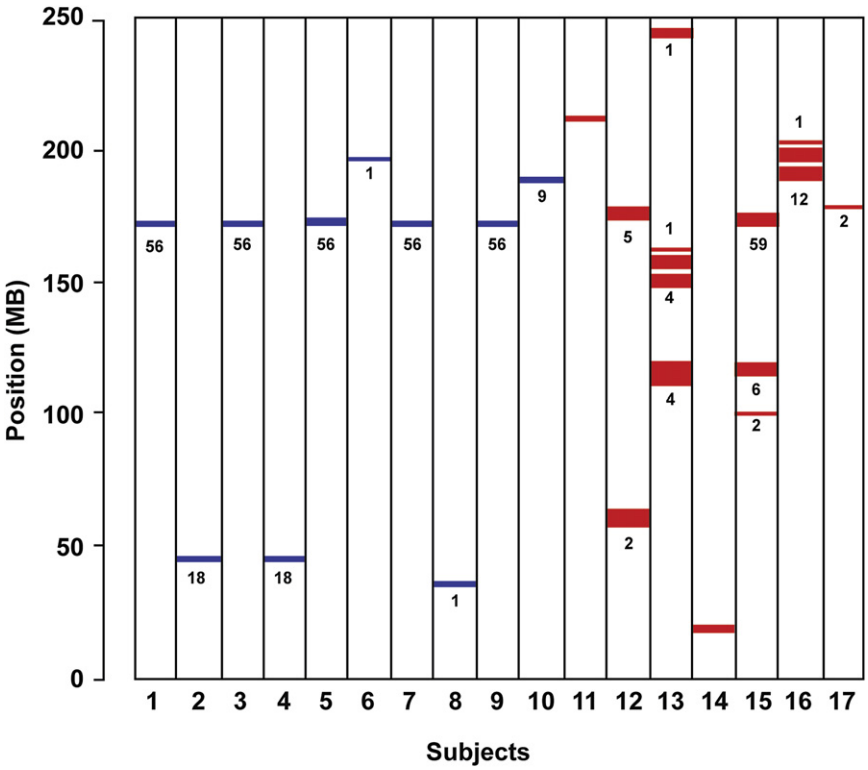


Figure 7. Size and Location of ROHs on Chromosome 1, Comparing Half Orcadians and Offspring of Cousins

ROHs measuring ≥ 1.5 Mb in ten half Orcadians are shown in blue, and those of seven offspring of 1st–3rd cousins are shown in red. The numbers shown below each colored segment are the numbers of overlapping ROHs in the Scottish control sample.

to refine the methodology, particularly in relation to the most appropriate length threshold for defining ROHs—but, rather, to outline a broad approach and highlight issues for future consideration. Equally, a detailed evaluation of alternative methods is beyond the scope of this paper; however, we have made some preliminary comparisons with two of the measures, F_{plink} and multilocal heterozygosity (MLH). Both correlate strongly with F_{roh}. Whereas 1 − MLH is a measure of genome-wide homozygosity[54] with no attempt to distinguish loci that are homozygous because of IBD and loci that are homozygous by chance, F_{plink}[46] uses expected genome heterozygosity to control for homozygosity by chance. Carothers et al.[20] have proposed another measure of autozygosity, which uses locus-specific heterozygosity to give more weight to polymorphic loci that are homozygous. Unlike our approach, all three methods are single-point approaches and do not exploit the nature of autozygosity that comes in runs or tracts. Another drawback of F_{plink} and the method proposed by Carothers et al. is that they require estimation of population allele frequencies, a nontrivial problem in many populations.[62] Leutenegger et al.[22] have also proposed a multipoint approach to autozygosity inference. Their method uses a hidden Markov model that requires that markers are in linkage equilibrium. Hence, it is computationally more complex to deal with extremely dense SNP maps, because LD needs to be taken into account or a subset of SNPs in low LD needs to be selected. Both of these are subject to ongoing research. The method is, on the other hand, very well suited for dense microsatellite maps or mixed microsatellite-SNP maps.[28]

F_{roh} differs from all these approaches in that it is based on the assumption that ROHs are a signature of autozygosity (Figure 1), which might be the result of recent parental relatedness but equally might be of much more ancient origin. This is clearly illustrated by our half-Orcadian population, whose parents are known to be unrelated and who, therefore, have inherited no IBD alleles for at least five and probably 10–12 generations. We show, however, that on average, half Orcadians have a total of 6 Mb worth of ROHs measuring longer than 1.5 Mb (0.2% of the autosomes). In the two nonisolate populations studied, the comparable statistics are 7.25 Mb (0.3% of autosomes), in the Scottish population, and 8.3 Mb (0.3%), in the CEU population (Figure 6).

Consistent with the findings of other studies,[56,59] we have shown that these shorter ROHs are almost invariably common but not universal in the population, occurring in both a Scottish control group (Figure 7) and an outbred non-Hispanic European American population.[56] Common ROHs are a source of individual genetic variation that might play a causal role in common complex disease and that, therefore, merit further exploration as risk factors in their own right.[56] We feel that it is also entirely appropriate to count them in our F_{roh} statistic for the purposes of investigating the effect of genome-wide homozygosity on quantitative disease or disease-related traits. For this purpose, we suggest a minimum-length threshold of 0.5 Mb, because this is the limit of resolution possible with a 300,000-SNP genome-wide scan and is also considerably longer than most stretches of LD.[48–51] There is, though, clearly potential for exploration of the prevalence and distribution of even-shorter ROHs with the use of data sets with more densely spaced markers.

When the research aim is to use homozygosity mapping to identify the variants causing rare recessive diseases, F_{roh} can be modified in order to reflect only the effects of recent parental relatedness. Our analysis of the genomic location of ROHs shows that many of the most common ROHs are equally present in the offspring of both related and unrelated parents (see Table 2 and Figure 7). We propose that

F_{roh} could be modified by identification of such common ROHs and removal of them from both the numerator and the denominator, thus reducing the risk of false negatives. An alternative approach would be to set a higher minimum-length threshold, for example, 5 Mb (see Table 1 and Figure 5), but this would have the effect of underestimating the effects of recent parental relatedness by failure to count any shorter ROHs of recent origin, while still not totally eliminating longer, common ROHs.

We have shown here that ROHs measuring 1.5 Mb and longer can be used to distinguish between populations with different histories of isolation. ROHs also distinguish effectively between individuals with different degrees of parental relatedness in their ancestry. This approach is simple, observational, and based on sound theoretical justification. Although our study is based on Illumina data, this method is generally applicable, and we see no reason why it could not be used with data generated on other platforms. With some refinement, F_{roh} has potential as a measure of individual genome-wide autozygosity for comparison to phenotype. The essential challenge in any attempt to estimate individual autozygosity from genomic data is to set a limit distinguishing autozygous from merely homozygous genotypes. Single-point methodologies based on estimation of population allele frequencies implicitly use time as a limit but face the serious drawback of requiring allele-frequency data for a founder or reference population. Our multipoint approach, which exploits the potential of ROHs as a measure of autozygosity, uses ROH length as a limit. Here, we have described how F_{roh} is affected by the length threshold used and by the inclusion of common ROHs. The next challenge is to establish the optimum-length threshold and determine to what extent F_{roh} should be modified with reference to the prevalence of common ROHs. These issues are the subject of ongoing research, involving the simulation of high-density genotype data by gene dropping fully phased Hap300 data down representative pedigrees. Work is also in progress to apply this approach to data sets from highly consanguineous populations and, in particular, to investigate whether the F_{roh} length cutoff used here is universally applicable. Common, shorter ROHs also merit further investigation as a risk factor in common complex disease and will have utility in narrowing down genomic regions in the search for functional genetic variants.[56] The availability of denser genome-wide scans with 1 million or more SNPs will facilitate more reliable identification and enumeration of shorter ROHs, and the use of these large data sets in different populations will improve understanding of the frequency of common ROHs and how these differ among populations.

Acknowledgments

We thank the people of Orkney; Lorraine Anderson and the research nurse team in Orkney; Rosa Bisset, Kay Lindsay, Gail Crosbie, and the administrative team at Public Health Sciences, University of Edinburgh; Ruby McMenemy, Sam Harcus, George Gray, Orkney Library and Archive, and the Orkney Family History Society for help with reconstructing pedigrees; Graeme Grimes, Colin Semple, Sarfraz Mohammed, Dave du Feu, Craig Nicol, and Lesley McGoohan for IT support; and Evi Theodoratu for advice relating to the Scottish data set. ORCADES is supported by the Chief Scientist Office of the Scottish Executive, the MRC Human Genetics Unit, the Royal Society, the Edinburgh Wellcome Trust Clinical Research Facility, and the European Union Framework Program 6. Ruth McQuillan is supported by a University of Edinburgh College of Medicine and Veterinary Medicine Ph.D. studentship. Rehab Abdel-Rahman is supported by a fund from the Supreme Council of Egyptian Universities. Igor Rudan is supported by grant no. 108-1080315-0302, Branka Janicijevic by grant no. 196-1962766-2763, Nina Smolej-Narancic by grant no. 196-1962766-2747, and Pavao Rudan by grant no. 196-1962766-2751 of the Croatian Ministry of Science, Education and Sport.

Received: June 29, 2008
Revised: August 12, 2008
Accepted: August 13, 2008
Published online: August 28, 2008

Web Resources

The URLs for data presented herein are as follows:

PLINK, http://pngu.mgh.harvard.edu/purcell/plink/
UCSC Genome Browser, http://genome.ucsc.edu

References

1. Keller, L., and Waller, D. (2002). Inbreeding effects in wild populations. Trends Ecol. Evol. *17*, 230–241.
2. Charlesworth, B., and Charlesworth, D. (1999). The genetic basis of inbreeding depression. Genet. Res. *74*, 329–340.
3. Bittles, A.H. (2003). Consanguineous marriage and childhood health. Dev. Med. Child Neurol. *45*, 571–576.
4. Khlat, M., and Khoury, M. (1991). Inbreeding and diseases: demographic, genetic, and epidemiologic perspectives. Epidemiol. Rev. *13*, 28–41.
5. Modell, B., and Darr, A. (2002). Science and society: genetic counselling and customary consanguineous marriage. Nat. Rev. Genet. *3*, 225–229.
6. Bonifati, V., Rizzu, P., van Baren, M.J., Schaap, O., Breedveld, G.J., Krieger, E., Dekker, M.C., Squitieri, F., Ibanez, P., Joosse, M., et al. (2003). Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. Science *299*, 256–259.
7. van Duijn, C.M., Dekker, M.C., Bonifati, V., Galjaard, R.J., Houwing-Duistermaat, J.J., Snijders, P.J., Testers, L., Breedveld, G.J., Horstink, M., Sandkuijl, L.A., et al. (2001). Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36. Am. J. Hum. Genet. *69*, 629–634.
8. Ewald, H., Wikman, F.P., Teruel, B.M., Buttenschon, H.N., Torralba, M., Als, T.D., El Daoud, A., Flint, T.J., Jorgensen, T.H., Blanco, L., et al. (2005). A genome-wide search for risk genes using homozygosity mapping and microarrays with 1,494 single-nucleotide polymorphisms in 22 eastern Cuban families with bipolar disorder. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *133*, 25–30.

9. Mani, A., Meraji, S.M., Houshyar, R., Radhakrishnan, J., Mani, A., Ahangar, M., Rezaie, T.M., Taghavinejad, M.A., Broumand, B., Zhao, H., et al. (2002). Finding genetic contributions to sporadic disease: a recessive locus at 12q24 commonly contributes to patent ductus arteriosus. Proc. Natl. Acad. Sci. USA *99*, 15054–15059.
10. Rudan, I., Smolej-Narancic, N., Campbell, H., Carothers, A., Wright, A., Janicijevic, B., and Rudan, P. (2003). Inbreeding and the genetic complexity of human hypertension. Genetics *163*, 1011–1021.
11. Krieger, H. (1969). Inbreeding effects on metrical traits in Northeastern Brazil. Am. J. Hum. Genet. *21*, 537–546.
12. Martin, A.O., Kurczynski, T.W., and Steinberg, A.G. (1973). Familial studies of medical and anthropometric variables in a human isolate. Am. J. Hum. Genet. *25*, 581–593.
13. Hurwich, B.J., and Nubani, N. (1978). Blood pressures in a highly inbred community–Abu Ghosh, Israel. 1. Original survey. Isr. J. Med. Sci. *14*, 962–969.
14. Halberstein, R.A. (1999). Blood pressure in the Caribbean. Hum. Biol. *71*, 659–684.
15. Campbell, H., Carothers, A.D., Rudan, I., Hayward, C., Biloglav, Z., Barac, L., Pericic, M., Janicijevic, B., Smolej-Narancic, N., Polasek, O., et al. (2007). Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum. Mol. Genet. *16*, 233–241.
16. Saleh, E.A., Mahfouz, A.A., Tayel, K.Y., Naguib, M.K., and Bin-al-Shaikh, N.M. (2000). Hypertension and its determinants among primary-school children in Kuwait: an epidemiological study. East. Mediterr. Health J. *6*, 333–337.
17. Badaruddoza. (2004). Inbreeding effects on metrical phenotypes among north Indian children. Collegium Antropologicum. *28* (Suppl(2)), 311–319.
18. Hartl, D., and Clark, A.G. (1997). Principles of Population Genetics (Sunderland, MA: Sinauer Associates).
19. Wright, S. (1922). Coefficients of inbreeding and relationship. Am. Nat. *56*, 330–338.
20. Carothers, A.D., Rudan, I., Kolcic, I., Polasek, O., Hayward, C., Wright, A.F., Campbell, H., Teague, P., Hastie, N.D., and Weber, J.L. (2006). Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. Ann. Hum. Genet. *70*, 666–676.
21. Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. *35*, 131–155.
22. Leutenegger, A.L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E.A. (2003). Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. *73*, 516–523.
23. Woods, C.G., Valente, E.M., Bond, J., and Roberts, E. (2004). A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. J. Med. Genet. *41*, e101.
24. Liu, F., Elefante, S., van Duijn, C.M., and Aulchenko, Y.S. (2006). Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Ann. Hum. Genet. *70*, 965–970.
25. Smith, C. (1953). The detection of linkage in human genetics. J. Royal Stat. Soc. B. *15*, 153–192.
26. Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. *33* (*Suppl*), 228–237.
27. Miano, M.G., Jacobson, S.G., Carothers, A., Hanson, I., Teague, P., Lovell, J., Cideciyan, A.V., Haider, N., Stone, E.M., Sheffield, V.C., et al. (2000). Pitfalls in homozygosity mapping. Am. J. Hum. Genet. *67*, 1348–1351.
28. Leutenegger, A.L., Labalme, A., Genin, E., Toutain, A., Steichen, E., Clerget-Darpoux, F., and Edery, P. (2006). Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. Am. J. Hum. Genet. *79*, 62–66.
29. Shami, S.A., Qaisar, R., and Bittles, A.H. (1991). Consanguinity and adult morbidity in Pakistan. Lancet *338*, 954.
30. Puzyrev, V.P., Lemza, S.V., Nazarenko, L.P., and Panphilov, V.I. (1992). Influence of genetic and demographic factors on etiology and pathogenesis of chronic disease in north Siberian aborigines. Arctic Med. Res. *51*, 136–142.
31. Ismail, J., Jafar, T.H., Jafary, F.H., White, F., Faruqui, A.M., and Chaturvedi, N. (2004). Risk factors for non-fatal myocardial infarction in young South Asian adults. Heart *90*, 259–263.
32. Simpson, J.L., Martin, A.O., Elias, S., Sarto, G.E., and Dunn, J.K. (1981). Cancers of the breast and female genital system: search for recessive genetic factors through analysis of human isolate. Am. J. Obstet. Gynecol. *141*, 629–636.
33. Lebel, R.R., and Gallagher, W.B. (1989). Wisconsin consanguinity studies. II: Familial adenocarcinomatosis. Am. J. Med. Genet. *33*, 1–6.
34. Rudan, I. (1999). Inbreeding and cancer incidence in human isolates. Hum. Biol. *71*, 173–187.
35. Carlson, C.S., Eberle, M.A., Kruglyak, L., and Nickerson, D.A. (2004). Mapping complex disease loci in whole-genome association studies. Nature *429*, 446–452.
36. Freimer, N., and Sabatti, C. (2004). The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat. Genet. *36*, 1045–1051.
37. Wright, A., Charlesworth, B., Rudan, I., Carothers, A., and Campbell, H. (2003). A polygenic basis for late-onset disease. Trends Genet. *19*, 97–106.
38. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.
39. Broman, K.W., and Weber, J.L. (1999). Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am. J. Hum. Genet. *65*, 1493–1500.
40. Tenesa, A., Farrington, S.M., Prendergast, J.G., Porteous, M.E., Walker, M., Haq, N., Barnetson, R.A., Theodoratou, E., Cetnarskyj, R., Cartwright, N., et al. (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat. Genet. *40*, 631–637.
41. The International HapMap Project. Nature *426*, 789–796.
42. Brennan, E.R., and Relethford, J.H. (1983). Temporal variation in the mating structure of Sanday, Orkney Islands. Ann. Hum. Biol. *10*, 265–280.
43. Boyce, A.J., Holdsworth, V.M.L., and Brothwell, D. (1973). Demographic and genetic studies in the Orkney islands. In Genetic Variation in Britain, D.F. Roberts and E. Sunderland, eds. (London: Taylor and Francis).
44. Vitart, V., Rudan, I., Hayward, C., Gray, N.K., Floyd, J., Palmer, C.N., Knott, S.A., Kolcic, I., Polasek, O., Graessler, J., et al. (2008). SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. Nat. Genet. *40*, 437–442.

45. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
46. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
47. Department, R.A., ed. (2004). *Scottish Executive Urban Rural Classification 2003–2004*, E.a.
48. Abecasis, G.R., Ghosh, D., and Nichols, T.E. (2005). Linkage disequilibrium: ancient history drives the new genetics. Hum. Hered. *59*, 118–124.
49. Wall, J.D., and Pritchard, J.K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. *4*, 587–597.
50. Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A., et al. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. Am. J. Hum. Genet. *68*, 191–197.
51. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.
52. Marioni, J.C., Thorne, N.P., Valsesia, A., Fitzgerald, T., Redon, R., Fiegler, H., Andrews, T.D., Stranger, B.E., Lynch, A.G., Dermitzakis, E.T., et al. (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol. *8*, R228.
53. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. *35*, 2013–2025.
54. Charpentier, M., Setchell, J.M., Prugnolle, F., Knapp, L.A., Wickings, E.J., Peignot, P., and Hossaert-McKey, M. (2005). Genetic diversity and reproductive success in mandrills (Mandrillus sphinx). Proc. Natl. Acad. Sci. USA *102*, 16723–16728.
55. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
56. Lencz, T., Lambert, C., DeRosse, P., Burdick, K.E., Morgan, T.V., Kane, J.M., Kucherlapati, R., and Malhotra, A.K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc. Natl. Acad. Sci. USA *104*, 19942–19947.
57. Simon-Sanchez, J., Scholz, S., Fung, H.C., Matarin, M., Hernandez, D., Gibbs, J.R., Britton, A., de Vrieze, F.W., Peckham, E., Gwinn-Hardy, K., et al. (2007). Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum. Mol. Genet. *16*, 1–14.
58. Gibson, J., Morton, N.E., and Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. Hum. Mol. Genet. *15*, 789–795.
59. Curtis, D., Vine, A.E., and Knight, J. (2008). Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann. Hum. Genet. *72*, 261–278.
60. Li, L.H., Ho, S.F., Chen, C.H., Wei, C.Y., Wong, W.C., Li, L.Y., Hung, S.I., Chung, W.H., Pan, W.H., Lee, M.T., et al. (2006). Long contiguous stretches of homozygosity in the human genome. Hum. Mutat. *27*, 1115–1121.
61. Vitart, V., Carothers, A.D., Hayward, C., Teague, P., Hastie, N.D., Campbell, H., and Wright, A.F. (2005). Increased level of linkage disequilibrium in rural compared with urban communities: a factor to consider in association-study design. Am. J. Hum. Genet. *76*, 763–772.
62. Hoffman, J.I., Boyd, I.L., and Amos, W. (2004). Exploring the relationship between parental relatedness and male reproductive success in the Antarctic fur seal Arctocephalus gazella. Evolution Int. J. Org. Evolution *58*, 2087–2099.