The Genomic Landscape of the Old Order Amish

by

Cristopher V. Van Hout

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Julie A. Douglas, Chair
Professor David T. Burke
Assistant Professor Sebastian K. Zoellner
Professor Jeffrey C. Long, University of Albuquerque, NM

Acknowledgements

First off, I recognize the effort and patience of my research mentor Julie Douglas; Thank you for teaching from the heart. I thank the Douglas lab members, Kaanan Shah, Jim MacDonald, Marie-Hélène Roy-Gagnon, Connor Sandefur, and Al Levin. I thank my thesis committee Dave Burke, Sebastian Zöllner, and Jeff Long for making the most of many teachable moments.

Also, I thank our collaborators in the Amish research studies, Pat Peyser in the Epidemiology department at the University of Michgian, Alan Shuldiner and Brackie Mitchell at the University of Maryland Baltimore.

In the Human Genetics department, I thank Janet Miller and Karen Grahl for helping me to navigate a sea of red tape. I thank Didi Robins and Sally Camper for their sincerity and kindness, and particularly Charlie Sing for putting the philosophy in my PhD.

To friends and colleagues at and around the University of Michigan, thank you; Tim Connallon and Akane Uesugi, Michelle Gornick, Jake Higgins and Alicia Levesque, Nicole Scott, Ali Shojaie, Mara Steinkamp, Stillian and Emily Stoev, and Kevin Vannella.

In particular, I thank my family, aunts Cheryl Mader, Janie Moll, Anna Ewanowski, and a small army of cousins for being amazing. I especially am grateful for the support of my sister Val Schroeder and my mother Mary Van Hout. Finally, I thank Randi Prieve for her encouragement and companionship; you have been my silver cord.

Table of Contents

List of Tables

Abstract

The Genomic Landscape of the Old Order Amish

by

Cristopher V. Van Hout

Chair: Julie A. Douglas

The Old Order Amish (OOA) of Lancaster County Pennsylvania are a population isolate with a census size of ~35,000 individuals who descended from ~200 immigrants from Western Europe in the early 1700s. They have a long history of participation in genetic studies, for which their genealogical records and simple lifestyle offer substantial research advantages. However, their demographic history has altered their genomic landscape relative to their European counterparts. Knowledge of this landscape is critical to the design, execution, and interpretation of genetic studies in the OOA. In this dissertation, I evaluate the consequences of population bottleneck and genetic drift on the empirical and/or expected distribution of 1) linkage disequilibrium (LD) for common variants, 2) rare variation (with a focus on the implications for imputation accuracy using an external population) and 3) genomic estimates of inbreeding in the OOA.

Using a high-density Single Nucleotide Polymorphism (SNP) map, I compare LD between OOA individuals and a reference population of European ancestry (HapMap CEU). For common SNPs (Minor Allele Frequency (MAF) ≥ 0.05), allele frequencies and LD profiles were similar between the OOA and CEU. Thus, public resources

constructed from CEU data are appropriate for analyses of common genetic variation in the OOA.

To assess the portability of deep sequencing resources, e.g., 1000 Genomes Project, for rare SNPs (MAF<0.05), I evaluate (via simulation and small-scale empirical study) the impact of using CEU versus OOA haplotype reference panels on imputation accuracy in the OOA. My results establish likely lower and upper bounds (0.50 and 0.75, respectively) of imputation accuracy for rare SNPs using 1000 Genomes Project-like resources in the OOA.

Finally, using a subset of SNPs from the high-density map above, I estimate genomic inbreeding coefficients and compare them inbreeding conditional on the OOA pedigree, and describe the distribution of autozygous segments in the study participants. I observed strong agreement between genomic- and pedigree-based estimates, with a mean inbreeding coefficient of ~0.035, approximately the offspring of half 1^{st} cousins. Furthermore, I establish that approximately 92% of the inbreeding in the OOA pedigree is due to inbreeding loops more distant than offspring of 2^{nd} cousins.

Chapter 1. Introduction

1.1 Founder populations and isolates

Founder populations, groups of people who are descended from substantially fewer number of individuals compared to cosmopolitan populations, have been popular populations of inference for disease and quantitative trait gene mapping [Ober and Cox 1998]. Isolated populations are reproductively distinct from other populations, and often characterized by a low rate of gene flow from other populations. Isolated and founder populations differ considerably in the number of founders and extent of isolation. For example, Finns [Hastbacka, et al. 1992; Palo, et al. 2009; Peltonen, et al. 1999] and Ashkenazi Jews [Bray, et al. 2010; Risch, et al. 1995] have large populations sizes, have been separated for relatively long time period, and have experienced substantial gene-flow with other groups relative to the Hutterites [Ober, et al. 2001] and the Old Order Amish (OOA) [Strauss and Puffenberger 2009] which is the population of inference of this dissertation. Each of these isolates has a distinct population history, which is often not accurately characterized, but uniquely shapes its genetic features. Thus, it is necessary to empirically evaluate the genomic landscape on a population-specific basis.

1.2 Population demography of the Old Order Amish

The Old Order Amish of Lancaster County Pennsylvania are a population isolate descended from an initial founding population of approximately 200 individuals of

Northern and Western European ancestry in the early 1700s [Lee, et al. 2010; McKusick, et al. 1964]. Since then, the Amish have grown to a census population size of approximately 35,000 individuals [Lee, et al. 2010]. The relationships between individuals over approximately 15 to 20 generations of population growth are documented by the OOA in The Fisher Book [Beiler 1988], which has been organized into the Anabaptist Genealogy DataBase (AGDB) [Agarwala, et al. 2001].

1.3 Genetic epidemiology in the Old Order Amish

The deep OOA genealogy has proved to be extremely useful for studies of disease and quantitative trait phenotypes in the OOA. Starting in the mid 1960s, physician geneticists initiated studies of inborn errors of metabolism in the OOA (for review, see [Strauss and Puffenberger 2009]). More recently, genome-wide linkage and association analyses have been conducted in the OOA for cardiovascular phenotypes [Mitchell, et al. 2008; Roy-Gagnon, et al. 2008], diabetes [Hsueh, et al. 2000], coronary artery disease [Post, et al. 2007], circulating lipid profiles [Pollin, et al. 2008a], and mammographic breast density [Douglas, et al. 2008].

The unique demographic history of the OOA has altered the genomic landscape, which must be taken into account in the design, execution, and interpretation of genetic epidemiological studies. It is in this capacity that I hope the research in this dissertation will inform ongoing studies in the OOA and guide future studies in this and other populations.

1.4 Organization of this dissertation

I examine the extent to which the demographic history has impacted the genomic landscape of the OOA in three projects: the effect on common and rare Single Nucleotide Polymorphisms (SNPs), the impact on rare variation, and the extent of inbreeding.

In Chapter 2, for common Single Nucleotide Polymorphisms (Minor Allele Frequency (MAF) > 0.05), I compared allele frequencies and linkage disequilibrium (LD) profiles between the OOA and the HapMap CEU. In the course of the analysis of genetic epidemiological studies, it is common to refer to public databases with densely genotyped individuals, such as the HapMap, to assess whether genetic markers of interest (regions of linkage, or SNPs with low p-values in GWAS) implicate obvious candidate genes. This essential interpretation step relies on similar patterns of LD between the population of interest and the reference population. Additionally, in the context of the common variant-common disease hypothesis, a demonstration that most common variants are indeed shared between the OOA and more cosmopolitan populations such as the CEU at least allows for the possibility that findings in the OOA may generalize to other populations. Prior to my publication of this work [Van Hout, et al. 2010], a genome-wide evaluation of allele frequency patterns and LD profiles in the OOA had not been published.

In Chapter 3, in contrast to common SNPs, rare SNPs (MAF < 0.05) could differ substantially in frequency between the OOA and CEU. Since we are increasingly able to measure rare SNPs, the evaluation of the contribution of rare variants to variation in disease phenotypes and quantitative traits is now an attainable goal. Imputing genotypes, i.e., using relatively few genotypes to identify the underlying haplotype from a more

densely typed or deeply sequenced reference panel, has attracted much attention as one approach for indirectly measuring genetic variation. This strategy relies on shared haplotype structure via shared ancestry between the haplotype reference and the study participants. Thus, in Chapter 3, I explore how population demography may have affected rare variation in the OOA in comparison to the CEU and evaluate imputation as a strategy for measuring rare variation in the OOA.

In Chapter 4, I compare estimates of inbreeding from genomic data to expectation conditional on pedigree information. From simulated genotypes of the offspring of 1st cousins, I show that the genomic-based estimator of inbreeding is better able to capture true inbreeding versus expectation conditional on the pedigree. Additionally, I estimate a locus-specific posterior probability of autozygosity to characterize the number and lengths of segments that are likely two alleles identical by descent in the OOA. Finally, I evaluate the extent to which inbreeding in the OOA is due to recent inbreeding loops, i.e., the offspring of 2nd cousins or closer.

In Chapter 5, I summarize the implications of this dissertation for genetic epidemiological and population genetic studies in the Old Order Amish and consider the applicability of these findings to other population isolates. Additionally, I discuss potential future research projects as extensions of the ideas developed in the course of this dissertation.

1.5 References

Agarwala R, Schaffer AA, Tomlin JF. 2001. Towards a complete North American Anabaptist Genealogy II: analysis of inbreeding. Hum Biol 73(4):533-45.

Beiler K, editor. 1988. Fisher Family History. Third ed: Eby's Quality Printing, PA.

Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST. 2010. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. Proc Natl Acad Sci U S A 107(37):16222-7.

Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2(3):204-11.

Hsueh WC, Mitchell BD, Aburomia R, Pollin T, Sakul H, Gelder Ehm M, Michelsen BK, Wagner MJ, St Jean PL, Knowler WC and others. 2000. Diabetes in the Old Order Amish: characterization and heritability analysis of the Amish Family Diabetes Study. Diabetes Care 23(5):595-601.

Lee WJ, Pollin TI, O'Connell JR, Agarwala R, Schaffer AA. 2010. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. BMC Med Genet 11:68.

McKusick VA, Hostetler JA, Egeland JA. 1964. Genetic Studies of the Amish, Background and Potentialities. Bull Johns Hopkins Hosp 115:203-22.

Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, Jaquish C, Douglas JA, Roy-Gagnon MH, Sack P and others. 2008. The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am Heart J 155(5):823-8.

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

Ober C, Cox NJ. 1998. The genetics of asthma. Mapping genes for complex traits in founder populations. Clin Exp Allergy 28 Suppl 1:101-5; discussion 108-10.

Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. 2009. Genetic markers and population history: Finland revisited. Eur J Hum Genet 17(10):1336-46.

Peltonen L, Jalanko A, Varilo T. 1999. Molecular genetics of the Finnish disease heritage. Hum Mol Genet 8(10):1913-23.

Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, Horenstein RB, Post W, McLenithan JC, Bielak LF, Peyser PA and others. 2008. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 322(5908):1702-5.

Post W, Bielak LF, Ryan KA, Cheng YC, Shen H, Rumberger JA, Sheedy PF, 2nd, Shuldiner AR, Peyser PA, Mitchell BD. 2007. Determinants of coronary artery and aortic calcification in the Old Order Amish. Circulation 115(6):717-24.

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, Breakefield X, Bressman S. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9(2):152-9.

Roy-Gagnon MH, Weir MR, Sorkin JD, Ryan KA, Sack PA, Hines S, Bielak LF, Peyser PA, Post W, Mitchell BD and others. 2008. Genetic influences on blood pressure response to the cold pressor test: results from the Heredity and Phenotype Intervention Heart Study. J Hypertens 26(4):729-36.

Strauss KA, Puffenberger EG. 2009. Genetics, medicine, and the Plain people. Annu Rev Genomics Hum Genet 10:513-36.

Chapter 2. Extent and Distribution of Linkage Disequilibrium in the Old Order Amish

2.1 Abstract

Knowledge of the extent and distribution of linkage disequilibrium (LD) is critical to the design and interpretation of gene mapping studies. Because the demographic history of each population varies and is often not accurately known, it is necessary to empirically evaluate LD on a population-specific basis. Here present the first genome-wide survey of LD in the Old Order Amish (OOA) of Lancaster County Pennsylvania, a closed population derived from a modest number of founders. Specifically, I present a comparison of LD between OOA individuals and U.S. Utah participants in the International HapMap project (abbreviated CEU) using a high-density single nucleotide polymorphism (SNP) map. Overall, the allele (and haplotype) frequency distributions and LD profiles were remarkably similar between these two populations. For example, the median absolute allele frequency difference for autosomal SNPs was 0.05, with an inter-quartile range of 0.02 to 0.09, and for autosomal SNPs 10-20 kb apart with common alleles (minor allele frequency ≥ 0.05), the linkage disequilibrium measure r^2 was at least 0.8 for 15% and 14% of SNP pairs in the OOA and CEU, respectively. Moreover, tag SNPs selected from the HapMap CEU sample captured a substantial portion of the common variation in the OOA (~88%) at $r^2 \geq 0.8$. These results suggest that the OOA and CEU may share similar LD profiles for other common but untyped SNPs. Thus, in the

context of the common variant-common disease hypothesis, genetic variants discovered in gene mapping studies in the OOA may generalize to other populations.

2.2 Introduction

Many genetic studies of complex traits and diseases are being conducted in population isolates, including the Old Order Amish (OOA) of Lancaster County Pennsylvania [Douglas, et al. 2008; Ginns, et al. 1998; Hsueh, et al. 2000; Mitchell, et al. 2001; Mitchell, et al. 2008; Post, et al. 2007; Streeten, et al. 2006; Wang, et al. 2009b; Y, et al. 2009]. Whether results from these studies will generalize to other populations is dependent (in part) on the similarity of allele frequencies and patterns of linkage disequilibrium between populations. To inform future genetic studies of the OOA and facilitate comparisons of findings with other populations, I conducted the first genome-wide survey of linkage disequilibrium in the OOA and compared our findings to the International HapMap project [Frazer, et al. 2007].

Most of the present-day OOA of Lancaster County are the descendants of approximately 200 individuals [Cross 1976] from central western Europe who immigrated to the United States in the early eighteenth century [Cross 1976; McKusick, et al. 1964]. Although recent data indicate that the differences in LD between isolated and cosmopolitan populations for common alleles are modest [Bonnen, et al. 2006; Service, et al. 2006], the uncertain but unique demographic history of the OOA necessitates empirical evaluation of LD.

2.3 Subjects and Methods

OOA study subjects were recruited and genotyped (n=861) in the course of the Heredity and Phenotype Intervention (HAPI) Heart study [Mitchell, et al. 2008], which was designed to identify gene-environment interactions influencing cardiovascular traits. Because many closely related individuals were deliberately ascertained, I used a simulated annealing algorithm [Douglas and Sandefur 2008] to select a set of minimally related individuals (30 men and 30 women) by minimizing the maximum pair-wise kinship coefficient of the set. The median [range] pair-wise kinship coefficient was 0.03 [0.01-0.04] for the set of 60 versus 0.03 [0.01-0.3] for the entire sample of 861. Notably, the maximum pair-wise kinship coefficient in the set of minimally related individuals was 0.04, i.e., no pair of individuals were closer than first cousins, which have a pair-wise kinship of 0.0625. For comparison with the OOA, I also utilized 30 men and 30 women (or 60 unrelated parents) from a U.S. Utah population with northern and western European ancestry (abbreviated CEU) in the International HapMap project [Frazer, et al. 2007].

2.3.1 Genotyping and QC Methods

DNA was extracted from whole blood by standard methods as described previously [Mitchell, et al. 2008]. The Affymetrix GeneChip® Human Mapping 500k Array Set was used for the comparison of LD patterns in both the OOA and CEU samples. Genotype calls were made using a Bayesian Robust Linear Model with Mahalanobis (BRLMM) distance classifier [Affymetrix 2006]. Genotype data for the CEU sample and corresponding annotation for the platform, including chromosome and

genomic positions for all SNPs on the array, were obtained from the Affymetrix website (www.affymetrix.com).

Individuals with >5% missing genotypes, and/or for men, >1% heterozygous genotypes on the X chromosome, were excluded. A subset of autosomal SNPs (2,068), which were selected to have high information content (minor allele frequency (MAF) ≥0.3), low pair-wise LD (maximum r^2 of 0.44), and coverage across all autosomes (average inter-marker spacing of 1.3 cM) in the OOA, were used to infer relationships using the maximum likelihood method implemented in Relpair [Epstein, et al. 2000]. I excluded individuals who had an inferred relationship that differed from the pedigree relationship with a likelihood ratio greater than 10^6. Based on these combined criteria, a total of 24 individuals (out of 861) were excluded from further analysis.

SNPs were required to satisfy the following quality control criteria in both samples: (1) ≤ 5% uncalled genotypes; (2) ≤5 and ≤1 Mendelian inconsistencies in OOA and CEU samples, respectively, using pedigree diagnostics as implemented in PedCheck [O'Connell and Weeks 1998]; and (3) Hardy Weinberg Equilibrium (HWE) p-value≥10^{-6} by Fisher's exact test [Wigginton, et al. 2005] as implemented in Haploview [Barrett, et al. 2005]. To assess genotyping accuracy, I used duplicate genotype data for 61 of the 861 OOA subjects for whom data from the Affymetrix Genome-Wide Human SNP Array 6.0 (overlap of 482,235 SNPs with Affymetrix GeneChip® Human Mapping 500k Array Set) were also available. Only SNPs with <2 duplicate inconsistencies were retained for analysis. Of the 500,447 genotypes that mapped to a single location in the human genome, 82,404 failed at least one QC measure in at least one sample. Those SNPs were removed, leaving a total of 409,071 autosomal [Table 2.1] and 8,972 X chromosome

[Table 2.2] SNPs. For the SNPs that passed our quality control criteria, the genotype consistency rate among 61 duplicate pairs was 99.4%.

2.3.2 Statistical Analyses

Fisher's exact test was used to compare allele frequency distributions between the OOA and CEU. For common SNPs (MAF≥0.05) on the same chromosome and within 10 Mb of each other, I used the Expectation-Maximization (EM) algorithm to obtain maximum likelihood estimates of two-SNP haplotype frequencies and measured pair-wise LD by the r^2 and D' statistics [Lewontin 1964]. Based on common SNPs, I also identified haplotype blocks in the CEU using an extension of the 4-gamete rule [Wang, et al. 2002] and estimated haplotype frequencies in both the CEU and OOA using the EM algorithm with a partition-ligation method [Qin, et al. 2002] for blocks with >10 SNPs as implemented in Haploview [Barrett, et al. 2005]. For each sample, I then calculated and compared the effective number of haplotypes in each block, i.e., $(\Sigma p_i^2)^{-1}$, where p_i is the frequency of the i^{th} haplotype in the block. As a measure of redundancy, I identified the number of SNPs (or proxies) that were in strong LD with each SNP at various thresholds of r^2 in each sample. To evaluate the extent to which SNPs selected to tag variation in the CEU capture common variation in the OOA, I selected common tag SNPs in the CEU using the greedy algorithm [Carlson, et al. 2004] implemented in Haploview [Barrett, et al. 2005] such that every unselected SNP had an $r^2 \geq 0.8$ with one or more selected SNPs. I then calculated r^2 between the tag SNPs and the remaining 'non-tagged' but typed SNPs in the OOA. Unless specified otherwise, all analyses were carried out using a combination of in-house R, Perl, and C programs.

2.4 Results

For the 418,043 SNPs that passed QC, mean heterozygosity was 0.26 and 0.27 for the autosomes in the OOA and CEU, respectively, and 0.23 and 0.24 for the X chromosome. The slightly lower heterozygosity in the OOA, in part, reflects the larger number of monomorphic SNPs in the OOA relative to the CEU, e.g., 68,869 versus 57,669 for the autosomes [Table 2.1]. For example, among the monomorphic SNPs in the OOA (n=16,869), 24% are polymorphic in the CEU, for which the median minor allele frequency is 0.017 with inter-quartile range of [0.008-0.025] and maximum 0.23. Among all SNPs that were polymorphic in at least one sample, the median absolute allele frequency difference was 0.05 for the autosomes and 0.07 for the X chromosome. At p-value$<10^{-6}$, OOA and CEU allele frequencies were significantly different for 799 autosomal and 137 X chromosome SNPs.

The percentage of SNP pairs within 10 Mb of each other and between which strong LD was observed was remarkably similar between the OOA and CEU for the autosomes [Table 2.3] and the X chromosome [Table 2.4]. For example, for autosomal SNPs at an inter-marker distance of <10 kb, no evidence of recombination (D'=1) was observed for 79% and 75% of SNP pairs, perfect LD (r^2=1) was observed for 20% and 19% of SNP pairs, and useful LD ($r^2 \geq 0.8$) was observed for 30% and 29% of SNP pairs in the OOA and CEU, respectively. Based on the CEU sample, I identified 58,097 autosomal haplotype blocks, with a median of 3 SNPs per block and an inter-quartile range of [3, 4]. Among all autosomal blocks, the median effective number of haplotypes (n_e) was 2.43 and 2.47 in the OOA and CEU, respectively, and the median of the differences in n_e (CEU minus OOA) per block was 0.04, with an inter-quartile range of -

0.2 to 0.3, suggesting modestly greater haplotype diversity in the CEU. A parallel analysis using haplotype blocks defined in the OOA did not qualitatively differ from the results based on blocks defined in the CEU.

Of common autosomal SNPs, 72% and 64% had at least one proxy at $r^2 \geq 0.8$ and 55% and 44% had at least one perfect proxy ($r^2=1$) in the OOA and CEU, respectively, indicating that fewer independent SNPs are required to represent variation in the OOA relative to the CEU. At $r^2 \geq 0.8$, 170,979 of 310,704 common SNPs in the CEU were selected as tag SNPs and captured ~88% of the 'non-tagged' SNPs in OOA, suggesting that SNPs selected to tag common variation in the CEU capture much of the same variation in the OOA. SNPs not captured by the CEU tag SNPs tended to be of lower minor allele frequency (data not shown). Results for the X chromosome were qualitatively similar.

2.5 Discussion

In general, I found a high degree of similarity in allele frequencies and LD patterns in the OOA and CEU samples. Allele frequencies were not significantly different between the OOA and CEU for >99% of SNPs. Of the SNPs that had significantly different allele frequencies, the proportion that were monomorphic was 1.7% and 0.9% in the OOA, and CEU, respectively. Based on common SNPs, which comprised 74% and 66% of autosomal SNPs in the OOA and CEU, respectively, the distribution and extent of LD were remarkably similar between these two samples. These data are consistent with previous theoretical predictions [Kruglyak 1999; Pritchard and Przeworski 2001] and recent empirical data [Bonnen, et al. 2006; Navarro, et al. 2009;

Service, et al. 2006; Thompson, et al. 2009], all of which point to modest differences in LD between isolated and cosmopolitan populations for common alleles. The situation for rare alleles, however, is likely to be different as has been demonstrated in applications of LD mapping for monogenic diseases and traits.

Demographic and historical information indicate that the OOA were founded relatively recently (~10 to 15 generations ago) by a modest number of individuals (several hundred) and then expanded rapidly to a current census population size exceeding 30,000 [Amish 2002]. Though the precise demographic details are unknown, it is apparent that the number of founders and rate of growth were sufficient and that the subsequent isolation of the OOA was too short for genetic drift and/or recombination to have meaningfully altered the common allele or haplotype frequency spectrum. Our recent study of variation on the Y chromosome supports these observations in that much of the diversity observed in non-isolated populations of similar ancestry is present in the OOA [Pollin, et al. 2008b]. It appears that inbreeding due to the finite population size of the OOA was also insufficient to meaningfully alter the allele frequency distribution or extent of LD. Based on the 60 OOA individuals included in our analyses, the average inbreeding coefficient F [Wright 1922] was 0.026 (range of 0.0003 to 0.046), which is too weak to generate substantial differences in LD relative to a non-isolated population [Hill and Robertson 1968].

Owing to similar allele frequencies and LD patterns in the OOA and CEU, CEU-derived tag SNPs performed well in capturing common variation in the OOA, consistent with previous studies in other samples of European ancestry, including those from isolated populations [Service, et al. 2007; Willer, et al. 2006]. These results suggest that

the OOA and CEU samples may also share similar LD profiles for other common but untyped SNPs. Thus, findings from gene mapping studies in the OOA may generalize to other populations in the context of the common variant-common disease hypothesis.

2.6 Acknowledgements

I gratefully acknowledge the Amish Research Clinic Staff, our Amish liaisons, and the Amish community, whose extraordinary support and cooperation made this study possible. I also thank Drs. Alejandro Schaffer and Richa Agarwala at the NIH/NCBI for providing the pedigree information and the Center for Inherited Disease Research (CIDR), NIH for providing duplicate genotypes from the Affymetrix Genome-Wide Human SNP Array 6.0. This study was supported in part by NIH grants U01 HL72515 and R01 CA122844.

The preceding chapter was published: Van Hout CV*****, Levin AM*****, Rampersaud E, Shen H, O'Connell J, Mitchell BD, Schuldiner AR, Douglas JA. Extent and Distribution of Linkage Disequilibrium in the Old Order Amish. Genetic Epidemiology, 2010 Feb;34(2):146-150. * The first two authors contributed equally to this work.

2.7 Tables

Table 2.1 Summary of autosomal SNPs

	OOA	CEU	Overlap
Total genotyped	489,922	489,922	489,922
>1 duplicate inconsistency[1]	51,459	NA	NA
>5% missing data[2]	50,085	16,896	8,973
Mendelian inconsistencies[2,3]	3,188	1,168	202
$p<10^{-6}$ for HWE test[4]	379	217	116
Passed QC filter[5]	415,440	472,851	409,071
Passed QC in both OOA and CEU			
Monomorphic[4]	68,869	57,669	52,467
Polymorphic[4]			
MAF≥0.05	297,605	310,704	287,476
MAF≥0.10	256,614	267,149	240,375
MAF≥0.20	182,941	189,133	161,062

OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency
Note: SNPs that failed a QC measure in either sample were excluded from further analysis, and SNPs with MAF≥0.05 passing QC in both samples (n=287,476) were used for LD analysis.
[1] Based on the 61 OOA individuals who were also genotyped on the Affymetrix 6.0 array; SNPs with more than one duplicated genotype discrepancy were excluded.
[2] Based on 837 OOA and 90 CEU individuals (30 trios).
[3] SNPs with >5 and >1 Mendelian inconsistencies in OOA and CEU, respectively.
[4] Based on 60 unrelated individuals (30 men and 30 women) from each sample.
[5] SNPs may fail QC in more than one way, so rows do not sum to the subtotal passing QC.

Table 2.2 Summary of X chromosome SNPs

	OOA	CEU	Overlap
Total genotyped	10,525	10,525	10,525
>1 duplicate inconsistency[1]	1,061	NA	NA
>5% missing data[2]	547	461	261
Mendelian inconsistencies[3,4]	44	246	10
$p<10^{-6}$ for HWE test[4]	0	0	0
Passed QC filter[5]	9,139	10,064	8,972
Passed QC in both OOA and CEU			
Monomorphic[4]	2,272	1,905	1,805
Polymorphic[4]			
MAF≥0.05	5,763	6,106	5,516
MAF≥0.10	4,971	5,376	4,449
MAF≥0.20	3,571	3,925	2,929

OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency
Note: SNPs that failed a QC measure in either sample were excluded from further analysis, and SNPs with MAF≥0.05 passing QC in both samples (n=5,516) were used for LD analysis.
[1] Based on the 61 OOA individuals who were also genotyped on the Affymetrix 5.0 array; SNPs with more than one duplicated genotype discrepancy were excluded.
[2] Based on 837 OOA and 90 CEU individuals (30 trios).
[3] SNPs with >5 and >1 Mendelian inconsistencies in OOA and CEU, respectively.
[4] Based on 60 unrelated individuals (30 men and 30 women) from each sample.
[5] SNPs may fail QC in more than one way, so rows do not sum to the subtotal passing QC.

Table 2.3 Linkage disequilibrium between autosomal SNPs

Percentage of autosomal SNP pairs[1] showing no evidence of recombination (D'=1), perfect LD (r^2=1), or where useful LD is observed (r^2≥0.8)

Inter-SNP distance (kb)	D'=1		r^2=1		r^2≥0.8	
	OOA	CEU	OOA	CEU	OOA	CEU
≤10	79	75	20	19	30	29
10-20	60	53	9	7	15	14
20-50	43	34	4	3	9	7
50-100	28	20	1	1	3	2
100-200	20	11	0	0	1	1
200-500	14	7	0	0	0	0
500-1,000	12	6	0	0	0	0
1,000-2,000	11	5	0	0	0	0
2,000-5,000	10	5	0	0	0	0
5,000-10,000	8	5	0	0	0	0

OOA = Old Order Amish (n=60)
CEU = U.S. Utah residents from HapMap (n=60)
[1] Restricted to SNPs with minor allele frequency ≥0.05 in both samples (n=287,476).

Table 2.4 Linkage disequilibrium between X chromosome SNPs

Percentage of X chromosome SNP pairs[1] showing no evidence of recombination (D'=1), perfect LD (r^2=1), or where useful LD is observed ($r^2 \geq 0.8$)

Inter-SNP distance (kb)	D'=1		r^2=1		$r^2 \geq 0.8$	
	OOA	CEU	OOA	OOA	CEU	OOA
≤10	88	85	39	35	51	49
10-20	72	64	23	19	34	31
20-50	60	48	12	9	21	18
50-100	44	31	6	3	11	10
100-200	31	19	3	1	6	4
200-500	22	11	1	0	2	1
500-1,000	18	7	0	0	0	0
1,000-2,000	17	7	0	0	0	0
2,000-5,000	15	7	0	0	0	0
5,000-10,000	13	7	0	0	0	0

OOA = Old Order Amish (n=60)
CEU = U.S. Utah residents from HapMap (n=60)
[1]Restricted to SNPs with minor allele frequency ≥0.05 in both samples (n=5,516).

2.8 References

Affymetrix. 2006. BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set. [http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf].

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-5.

Bonnen PE, Pe'er I, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly MJ, Reich DE and others. 2006. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat Genet 38(2):214-7.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74(1):106-20.

Cross HE. 1976. Population studies and the Old Order Amish. Nature 262(5563):17-20.

Douglas JA, Roy-Gagnon MH, Zhou C, Mitchell BD, Shuldiner AR, Chan HP, Helvie MA. 2008. Mammographic breast density--evidence for genetic correlations with established breast cancer risk factors. Cancer Epidemiol Biomarkers Prev 17(12):3509-16.

Douglas JA, Sandefur CI. 2008. PedMine--a simulated annealing algorithm to identify maximally unrelated individuals in population isolates. Bioinformatics 24(8):1106-8.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. Am J Hum Genet 67(5):1219-31.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61.

Ginns EI, St Jean P, Philibert RA, Galdzicka M, Damschroder-Williams P, Thiel B, Long RT, Ingraham LJ, Dalwaldi H, Murray MA and others. 1998. A genome-wide search for chromosomal loci linked to mental health wellness in relatives at high risk for bipolar affective disorder among the Old Order Amish. Proc Natl Acad Sci U S A 95(26):15531-6.

Hill WG, Robertson A. 1968. Linkage Disequilibrium in Finite Populations. Theoretical and Applied Genetics 38:226-231.

Hsueh WC, Mitchell BD, Aburomia R, Pollin T, Sakul H, Gelder Ehm M, Michelsen BK, Wagner MJ, St Jean PL, Knowler WC and others. 2000. Diabetes in the Old Order Amish: characterization and heritability analysis of the Amish Family Diabetes Study. Diabetes Care 23(5):595-601.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22(2):139-44.

Lancaster County Amish. 2002. Church Directory of the Lancaster County Amish: The Diary, Gordonville, PA.

Lewontin RC. 1964. The Interaction of Selection and Linkage. II. Optimum Models. Genetics 50:757-82.

McKusick VA, Hostetler JA, Egeland JA. 1964. Genetic Studies of the Amish, Background and Potentialities. Bull Johns Hopkins Hosp 115:203-22.
Mitchell BD, Hsueh WC, King TM, Pollin TI, Sorkin J, Agarwala R, Schaffer AA, Shuldiner AR. 2001. Heritability of life span in the Old Order Amish. Am J Med Genet 102(4):346-52.
Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, Jaquish C, Douglas JA, Roy-Gagnon MH, Sack P and others. 2008. The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am Heart J 155(5):823-8.
Navarro P, Vitart V, Hayward C, Tenesa A, Zgaga L, Juricic D, Polasek O, Hastie ND, Rudan I, Campbell H and others. 2009. Genetic comparison of a Croatian isolate and CEPH European Founders. Genetic Epidemiology (In this issue.).
O'Connell JR, Weeks DE. 1998. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63(1):259-66.
Pollin TI, McBride DJ, Agarwala R, Schaffer AA, Shuldiner AR, Mitchell BD, O'Connell JR. 2008. Investigations of the Y chromosome, male founder structure and YSTR mutation rates in the Old Order Amish. Hum Hered 65(2):91-104.
Post W, Bielak LF, Ryan KA, Cheng YC, Shen H, Rumberger JA, Sheedy PF 2nd, Shuldiner AR, Peyser PA, Mitchell BD. 2007. Determinants of coronary artery and aortic calcification in the Old Order Amish. Circulation 115(6):717-24.
Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1):1-14.
Qin ZS, Niu T, Liu JS. 2002. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 71(5):1242-7.
Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha JA and others. 2006. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38(5):556-60.
Service S, Sabatti C, Freimer N. 2007. Tag SNPs chosen from HapMap perform well in several population isolates. Genet Epidemiol 31(3):189-94.
Streeten EA, McBride DJ, Pollin TI, Ryan K, Shapiro J, Ott S, Mitchell BD, Shuldiner AR, O'Connell JR. 2006. Quantitative trait loci for BMD identified by autosome-wide linkage scan to chromosomes 7q and 21q in men from the Amish Family Osteoporosis Study. J Bone Miner Res 21(9):1433-42.
Thompson EE, Sun Y, Nicolae D, Ober C. 2009. Shades of gray: A comparison of linkage disequilibrium between Hutterites and Europeans. Genetic Epidemiology (In this issue.).
Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71(5):1227-34.
Wang Y, O'Connell JR, McArdle PF, Wade JB, Dorff SE, Shah SJ, Shi X, Pan L, Rampersaud E, Shen H and others. 2009. Whole-genome association study identifies STK39 as a hypertension susceptibility gene. PNAS 106(1):6.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76(5):887-93.
Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, Pruim R, Bark CW, Tsai YY, Pugh EW, Doheny KF and others. 2006. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. Genet Epidemiol 30(2):180-90.
Wright S. 1922. Coefficients of Inbreeding and Relationship. American Naturalist 56:330-338.

Chapter 3. Cataloging rare variants in the Old Order Amish - implications for imputation accuracy in isolated populations

3.1 Abstract

A substantial fraction of the genetic component of complex traits remains unexplained by the results of recent genome-wide association analyses of common SNPs, the so-called missing heritability problem. Thus, the contribution of rare variation to heritability is currently of great interest. Efforts such as the 1000 Genomes Project [Durbin, et al. 2010] are underway, in part, to catalog rare variants by deeply sequencing reference population panels.

For common SNPs (minor allele frequency ≥5%), I recently showed that the Old Order Amish (OOA) of Lancaster County, PA, an isolated population derived from a modest number of founders, and the HapMap CEU participants share similar allele frequencies and linkage disequilibrium profiles [Van Hout, et al. 2010]. Accordingly, I expect that reference panels like the CEU will adequately characterize common SNPs in the OOA. However, for rare SNPs, the OOA and CEU may differ considerably. Thus, in order to assess the portability of deep sequencing projects like the 1000 Genomes Project, I evaluated via simulation the impact of the population of origin of the haplotype reference panel (CEU versus OOA) on the imputation accuracy for rare SNPs in the OOA. In addition, I used CEU and OOA empirical genotypes on chromosome 22 to

impute rare SNPs in the OOA and the CEU using the 1000 Genomes Project low coverage Pilot sequence as the haplotype reference panel.

Using coalescent theory I simulated 100 megabases of sequence representative of the CEU and OOA, including 800 CEU-like and 800 OOA-like haplotypes to serve as reference panels and another 800 CEU-like and 800 OOA-like haplotypes to construct genotypes for pseudo-study participants. I masked ~95% of the study participants' genotypes and used the remaining 5% to impute the masked data from each reference panel. I characterized imputation accuracy by two measures: the coefficient of determination between the most likely imputed genotype and the true genotype, r^2, and the proportion of truly heterozygous genotypes that are imputed correctly.

As expected, based on simulations, for SNPs with MAF>5%, imputation accuracy as measured by r^2 was 93% and 96% based on the CEU-like and OOA-like reference panels, respectively. Similarly, for rare SNPs, 0.005<MAF<0.05, imputation accuracy was 75% and 86% based on the CEU-like and OOA-like reference panels, respectively. In the analysis using low coverage CEU data from the 1000 Genomes Pilot, imputation accuracy was lower. For example, for rare SNPs, 0.01<MAF<0.05, imputation accuracy as measured by r^2 was 0.50 in the OOA, consistent with the availability of fewer reference haplotypes in the low coverage 1000 Genomes Pilot data (120) compared to the simulated data (800).

3.2 Introduction

A substantial fraction of the prevalence of non-Mendelian disease is unexplained by recent genome-wide association studies (GWAS), the so-called missing heritability

problem [Manolio, et al. 2009]. Of the biological phenomenon that could contribute to the missing heritability, rare genetic variation is of considerable interest due, in part, to the development of genome-scale sequencing technology [Cirulli and Goldstein 2010]. However, some study strategies do not necessarily require direct observation of genetic variation. For example, using a much smaller sample of SNPs, a common strategy is to impute the unobserved variation by identifying underlying haplotype blocks from a deeply sequenced reference panel, where the haplotypes are shared due to common ancestry. In particular, the availability of deeply sequenced reference haplotype panels such as those of the 1000 Genomes Project [Durbin, et al. 2010] is expected to facilitate the imputation of sequence in diverse populations.

In previous work [Van Hout, et al. 2010], I compared allele frequencies and LD profiles for approximately 250,000 SNPs with minor allele frequency (MAF) greater than 5% in the OOA and the HapMap panel of Northern and Western European ancestry (CEU). I found that the OOA and CEU have similar LD profiles, so I predict high imputation accuracy for common SNPs in the OOA when using a CEU haplotype reference panel. However, for rare SNPs, expect that the OOA and CEU may differ considerably. Here, focus on the impact of the choice of reference panel on the imputation accuracy of rare SNPs, which define as between 0.5% and 5% MAF. Though empirical comparisons of the similarity of haplotype patterns have been carried out for a number of populations [Huang, et al. 2009], these studies have been limited to common SNPs. Moreover, for rare SNPs, populations that have been reproductively isolated may differ substantially from more cosmopolitan populations like those represented in the 1000 Genomes Project. Thus, the suitability of reference sequence, like those of the 1000

Genomes Project, for the purpose of imputation of rare SNPs is central to future studies of the contribution of rare variation to complex traits in population isolates, like the OOA. However, due to the lack of sequence data for a sufficient number of OOA individuals, I resort to simulating sequence data that is consistent with the known demographic history of the OOA. Specifically, I compared the expected performance of 1000 Genomes-like resources as a haplotype reference panel for imputation of rare SNPs in the Amish to a reference panel composed of OOA individuals, i.e., a population-specific panel. Additionally, using high density SNP data [Mitchell, et al. 2008], I evaluated imputation accuracy using the resources that are currently available.

3.3 Methods for the analysis of imputation accuracy using simulated data

To evaluate imputation accuracy, I simulated data representative of the expected release of phase one of the 1000 Genomes Project data which is expected to contain 400 individuals of European descent who would be appropriate to include in a haplotype reference panel for imputation of genotypes in the OOA. Though these 400 individuals are from four distinct ancestries, namely 100 individuals each from Utah with Northern and Western Europe ancestry (CEU), Italy (TSI), Britain (GBR), and Finland (FIN) (1000genomes.org), I assumed that all 400 individuals were CEU-like.

CEU-like haplotypes were simulated by a coalescent process by the method described by Kingman[Kingman 1982] and implemented by Hudson [Hudson 2002] and a parameterization that generates haplotypes with allele frequency spectra and linkage disequilibrium patterns that are consistent with those of HapMap CEU [Schaffner, et al. 2005]. Amish-like haplotypes were simulated by sampling 400 haplotypes from the

ancestral CEU-like population 20 generations in the past, after modeling exponential growth to an extant population size of 55,000. This parameterization reflects the demographic history of the OOA population, which is thought to have been founded by approximately 200 individuals in the early 1700s and expanded to a census size of approximately 35,000 individuals [Beiler 1988; Lee, et al. 2010]. The increase in extant population in the coalescent parameterization in comparison to the census size is intended to account for emigration from OOA community. Migration between the CEU-like population and the Amish-like population was assumed to be zero, consistent with the long term reproductive isolation and the negligible impact of gene flow from the Amish population to the CEU population. All other parameters were left at their default values, with a coalescent effective population size of the CEU-like population of 100,000, mutation rate per nucleotide per generation of $1.5x10^{-8}$, and a gene conversion rate of $4.5x10^{-9}$. Each simulated haplotype was one million nucleotides in length.

I generated three different configurations of simulated sequence. In each configuration, 800 haplotypes were chosen to form the haplotype reference panel, and 800 haplotypes were randomly paired to form the diploid genotypes of the 400 pseudo-study participants. First, 800 OOA-like and 800 CEU-like haplotypes were simulated from the same coalescent tree, representing a study strategy that uses an external, CEU-like reference panel consisting of 800 haplotypes to impute genotypes for 400 OOA pseudo-study participants. Second, 1600 OOA-like haplotypes were simulated, representing a study strategy that uses an internal, or study specific, reference panel consisting of 800 haplotypes to impute genotypes for 400 OOA pseudo-study participants. Third, for completeness, 1600 CEU-like haplotypes were simulated,

representing a study strategy that uses an external reference panel of 800 haplotypes for imputation for 400 CEU pseudo-study participants. One hundred independent replicates of each of the three configurations were generated.

For each replicate, 5% of non-monomorphic sites in the study participants were randomly selected as observed genotypes, while the remaining 95% of genotypes were masked. Uniformly masking 95% of SNPs resulted in an allele frequency spectrum for the observed SNPs that closely resembled the frequency spectrum on the Affymetrix 500k chip [Table 3.1]. The observed genotypes were used to impute the masked genotypes from the haplotype reference panel using the Markov chain haplotyping algorithm implemented in MaCH [Li, et al. 2009; Li, et al. 2010] with greedy scoring and 20 iterations of the Markov chain with sampling from all 800 haplotypes. The solution for each imputed genotype is specified as the most likely genotype, i.e., the genotype that was imputed most often across the iterations of the Markov chain.

I estimated imputation accuracy by two different measures: r^2, which is defined as the square of correlation between the true genotype and the most likely imputed genotype as estimated by an expectation maximization algorithm, and heterozygous agreement, abbreviated 'HetAgree', which is the proportion of truly heterozygous genotypes that were imputed correctly. To investigate how imputation accuracy differs by minor allele frequency, I computed the mean imputation accuracy for each measure within bins defined by the MAF in the haplotype reference panel. Because imputation accuracy decreases near the ends of the simulated haplotypes, due to reduced information in the observed SNPs (data not shown), genotypes within 200 kilobases of either end of

the simulated sequence were omitted from measures of accuracy to minimize the edge effects on imputation accuracy.

3.4 Methods for the analysis of imputation accuracy using empirical data

To evaluate imputation accuracy using empirical data, I limited the scope of analysis to chromosome 22 for simplicity. I constructed two analysis scenarios, both using empirical CEU sequence data as the haplotype reference panel to impute genotypes for 1) OOA and 2) CEU individuals. Specifically, the low density pilot project 1 draft of the 1000 Genomes Project, which consists of approximately 2-fold genome wide coverage for 60 unrelated CEU individuals, was used as the haplotype reference panel. The July 2010 draft of these data was downloaded from http://www.sph.umich.edu/csg/abecasis/MaCH/download/. This draft included 101,568 variable sites on chromosome 22, of which 70,572 were annotated with refSeq IDs. As samples for CEU and OOA individuals, I used genotypes from the Affymetrix 500k SNP chip for the same 60 unrelated CEU individuals provided by Affymetrix (http://www.affymetrix.com/support/technical/sample_data/500k_hapmap_genotype_data.affx) and 60 minimally related Amish individuals [Van Hout, et al. 2010]. Based on annotation information provided by Affymetrix, these data include 6,102 SNPs with refSeq IDs on chromosome 22, of which 5,100 were annotated with refSeq IDs in the 1000 Genomes project draft data.

I uniformly masked 5% of the 5,100 SNPs with MAF < 0.1 on chromosome 22 in the 60 CEU and OOA study participants and compared the masked genotypes to those imputed using the 1000 Genomes pilot data as the haplotype reference panel. Settings for

imputation using the MaCH algorithm and definitions of imputation accuracy were the same as those described for the simulated data. To remove the potential for differences in genotype strand orientation to falsely reduce imputation accuracy, SNPs in which the strand could not be unambiguously resolved, e.g., C/G or A/T SNPs, were omitted from the analysis. I repeated the masking and imputation of genotypes on chromosome 22 two hundred times.

3.5 Results

The mean imputation accuracies for the three configurations of study population and haplotype reference panel are given in Table 3.1. In general, imputation accuracy increased with minor allele frequency, and for rare SNPs, imputation accuracy was higher when the haplotype reference panel was drawn from the same population as the study population. As expected, for common SNPs in the OOA (MAF > 0.05 in the reference panel) imputation accuracy by the r^2 and HetAgree measures was high, 0. 93 and 0.97, respectively, using a CEU-like haplotype reference compared to 0.96 and 0.98, respectively, using the OOA-like haplotype reference. In contrast, for rare SNPs in the OOA (0.005 < MAF < 0.05 in the reference panel) r^2 and HetAgree were 0.75 and 0.82, respectively, using a CEU-like haplotype reference and 0.86 and 0.90, respectively, using an OOA-like haplotype reference. The inter-quartile range (IQR) for the r^2 measure of imputation accuracy in simulated haplotypes is given in [Table 3.2]. As expected, I observed greater variability in imputation accuracy for lower minor allele frequency SNPs.

I also classified the imputation errors by type using the CEU reference haplotypes to impute genotypes in the OOA. As expected, for rare SNPs ($0.005 < MAF \leq 0.05$), approximately 75% of SNPs that are imputed incorrectly are instances in which truly heterozygous genotypes are imputed incorrectly as major allele homozygous genotypes [Table 3.3]. The distribution of imputation errors was not qualitatively different for the other simulation scenarios [Data not shown].

Imputation accuracy using empirical 1000 Genomes Project low coverage pilot data as the haplotype reference for imputation of OOA and CEU genotypes was uniformly lower than in the simulated data [Table 3.4]. For SNPs in the OOA, $0.01 < MAF \leq 0.05$, imputation accuracy by the r^2 and HetAgree measures was 0.50 and 0.35 using a CEU-like haplotype reference. This difference is likely due, in part, to the decreased number of reference haplotypes available in the empirical data (n=120) compared to the simulations (n=800). Inter-quartile range (IQR) for the r^2 measure of empirical imputation error is given in Table 3.5.

3.6 Discussion and conclusions

By comparing an optimistic scenario using simulated data with simplifying assumptions of no genotyping or phasing error to a scenario using existing but limited resources (which are improving rapidly), I have established upper and lower bounds for the imputation accuracy that are likely to be observed in practice. Specifically, simulations imputing OOA-like genotypes using a CEU-like reference panel mimics a realistic study in which haplotypes for 400 individuals of European ancestry in the full Phase 1 1000 Genomes Project are available, for which the mean imputation accuracy for

rare SNPs ($0.005 < MAF \leq 0.05$) by the r^2 measure is expected to be approximately 0.75. Using 1000 Genomes pilot data for chromosome 22, consisting of haplotypes for only 60 CEU individuals to impute genotypes in the OOA, the mean imputation accuracy for rare SNPs was approximately 0.5. Because the r^2 is directly related to the sample size, this can be interpreted as requiring an approximate doubling of the number of study participants to retain equivalent power to detect an effect in a test of association.

I have shown that imputation accuracy of rare SNPs in the Old Order Amish is only marginally improved by using a population specific haplotype reference panel. For example, the r^2 measures of imputation accuracy for rare SNPs in the OOA-like population were 0.86 and 0.75 using a population-specific OOA-like haplotype reference and a CEU-like panel, respectively. These observations are consistent with the finding of an increase in imputation accuracy by the r^2 measure of approximately 4% for rare SNPs using population specific haplotype reference in the Finns [Surakka, et al. 2010]. Furthermore, these results suggest that studies using SNP data of a similar density to that of the Affymetrix 500k chip are likely to observe modest imputation accuracy for rare SNPs compared to studies using chips with a higher density. Specifically, the union of Illumina 1M and Affymetrix 6.0 arrays, i.e. HapMap 3, for imputation contain approximately 5 times the density of SNPs as the Affymetrix 500k chip. Studies using HapMap 3 to impute genotypes report substantially higher imputation accuracy in comparison to the present study [Altshuler, et al. 2010; Durbin, et al. 2010].

Through the simulated haplotypes in the OOA and CEU, it is possible to gain insight into the expected differences in the full allele frequency spectra between the two populations [Table 3.6]. For example, comparison of simulated allele frequencies in the

CEU and OOA suggest that most of the variation in the OOA is likely to be cataloged by a reference sample of 400 individuals of European ancestry as in Phase 1 of the 1000 Genomes Project. Approximately 1.8% of the variable sites (variable in the OOA or CEU) had MAF>0.005 in 800 OOA-like haplotypes but were monomorphic in the CEU-like haplotypes [Table 3.6]. Thus, even rare variation in the OOA is likely to be well represented in the 1000 Genomes Project data. Furthermore, though only a small proportion of alleles that are monomorphic in a sample of 400 CEU individuals are predicted to drift to MAF > 0.01 in the OOA, it is noteworthy at least one example of an allele that has drifted to substantially higher allele frequency in the OOA has been documented. Specifically, the G55T allele of the APOC3 gene, rs76353203, has MAF 0.05 in the OOA, but is monomorphic in a sample of 214 unrelated 'Caucasians' [Pollin, et al. 2008a]. It is likely that over the length of the genome, while the proportion may be small, the number of alleles that may have drifted to meaningfully higher frequency in the OOA could be large. In this context, studies of isolated populations may deliver substantially higher power to detect the contribution of these alleles to complex phenotypes, particularly given the increased power for the identifying rare sequence variation in large pedigrees.

Finally, the simplifying assumptions of the simulations, most notable error free sequence information, that have been made in the course of the current study represent limitations to the inference that should be drawn from these results. Further tuning of the coalescent parameters, including a parameterization that accounts for four distinct European populations, simulation of haplotypes longer than one megabase in length, and increasing the number of iterations of the imputation algorithm, could all result in

estimates of imputation accuracy that are closer to those of full scale empirical analyses. Limitations of the analysis of empirical data for chromosome 22 include the limited availability of haplotypes in the reference panel, i.e., 120 haplotypes vs. 800 Phase 1 1000 Genomes Project. Also, a more comprehensive analysis would include all available data, instead of focusing on chromosome 22. However, given computational constraints, and in the absence of empirical data, the development of upper and lower bounds for the accuracy of imputation for rare SNPs that are likely to be observed in a large scale study are a necessary step toward evaluating whether and how to integrate genotype imputation strategies into existing genetic epidemiological studies. Populations with similar demographic histories, such as the Hutterites [Thompson, et al. 2010], are likely to have similar imputation accuracies to those estimated in the Old Order Amish. Additionally, the simulation strategy that I implemented to evaluate the predicted performance of resources like the 1000 Genomes project could be adapted to model other populations.

3.7 Acknowledgements

I thank Drs. Braxton Mitchell and Alan Shuldiner, for their collaboration on the HAPI Heart Study. I acknowledge Matt Zawistowski and Sebastian Zollner for their helpful suggestions, particularly regarding the implementation of the imputation strategy. I gratefully acknowledge the Amish Research Clinic Staff, our Amish liaisons, and the Amish community, whose support and cooperation made this study possible. This research was supported by NIH grants T32 HG00040, U01 HL0201, and R01 (CA122844).

3.8 Tables

Table 3.1. Distribution of the number of SNPs (proportion of 1000 Genomes[1]) on chromosome 22 in different datasets

	1000 Genomes[1]	Affymetrix[2]	Uniform 5%[3]
$0 < MAF \leq 0.05$	31,460	699 (0.02)	1,573 (0.05)
$0.05 < MAF \leq 0.1$	15,433	551 (0.04)	772 (0.05)
$0.1 < MAF \leq 0.2$	18,340	937 (0.05)	917 (0.05)
$0.2 < MAF \leq 0.3$	13,633	834 (0.06)	682 (0.05)
$0.3 < MAF \leq 0.4$	11,204	700 (0.06)	560 (0.05)
$0.4 < MAF \leq 0.5$	10,980	657 (0.06)	549 (0.05)
$MAF > 0$	101,050	4,378 (0.04)	5,053 (0.05)

Counts (proportions) of SNPs for 60 unrelated CEU individuals from different datasets binned by MAF.
MAF = Minor Allele Frequency
[1] Genetic variants in the 1000 Genomes low coverage pilot, July 2010 annotation.
[2] Non-monomorphic SNPs that pass quality control measures (see text for details) on the Affymetrix 500k chip
[3] Five percent of the SNPs in the 1000 Genomes pilot data, i.e. the product of 0.05 and the number of SNPs in each MAF bin.

Table 3.2. Mean imputation accuracy in simulated sequence

Reference population[1]	CEU		OOA		CEU	
Study population[2]	OOA		OOA		CEU	
	r^2	HetAgree	r^2	HetAgree	r^2	HetAgree
0.005 <MAF≤ 0.01	0.52	0.67	0.77	0.81	0.64	0.66
0.01 <MAF≤ 0.025	0.77	0.76	0.89	0.87	0.79	0.76
0.025 <MAF≤ 0.05	0.87	0.83	0.93	0.91	0.85	0.81
0.05 <MAF≤ 0.1	0.93	0.88	0.95	0.93	0.91	0.86
0.1 <MAF≤ 0.5	0.97	0.93	0.98	0.96	0.96	0.92
0.005 <MAF≤ 0.05	0.75	0.82	0.86	0.90	0.73	0.80
0.05 <MAF≤ 0.5	0.93	0.97	0.96	0.98	0.92	0.96

[1] Ancestry of the haplotype reference panel
[2] Ancestry of the individuals for which genotypes are imputed
OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency in the reference population
r^2 = Squared correlation between the true genotype and the most likely imputed genotype
HetAgree = Proportion of truly heterozygous genotypes that were imputed correctly
The values reported are the means of measures of imputation accuracy for 100 replicates of simulated haplotypes one megabase in length for each reference/study configuration.
For each minor allele frequency bin, the minimum number of SNPs was 40,000.

Table 3.3. Inter-Quartile range for r^2 measure of imputation accuracy in simulated sequence

Reference population[1]	CEU		OOA		CEU	
Study population[2]	OOA		OOA		CEU	
	1stQ	3rdQ	1stQ	3rdQ	1stQ	3rdQ
0.005 <MAF≤ 0.01	0.42	0.97	0.71	0.99	0.40	0.97
0.01 <MAF≤ 0.025	0.60	0.98	0.81	0.99	0.59	0.98
0.025 <MAF≤ 0.05	0.76	0.99	0.88	1	0.71	0.98
0.05 <MAF≤ 0.1	0.85	0.99	0.93	1	0.84	0.99
0.1 <MAF≤ 0.5	0.94	1	0.96	1	0.93	1
0.005 <MAF≤ 0.05	0.57	0.98	0.80	0.99	0.54	0.98
0.05 <MAF≤ 0.5	0.92	1	0.96	1	0.91	1

[1] Ancestry of the haplotype reference panel
[2] Ancestry of the individuals for which genotypes are imputed
OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency in the reference population
r^2 = Squared correlation between the true genotype and the most likely imputed genotype
The values reported are the first quartile and third quartile of the distribution of the r^2 measure of imputation accuracy for 100 replicates of simulated haplotypes one megabase in length for each reference/study configuration. For each minor allele frequency bin, the minimum number of SNPs was 40,000.

Table 3.4. Distribution of imputation errors

Error type[1]	P(AA|Aa)	P(Aa|AA)	All Other
0.005<MAF ≤ 0.01	0.913	0.084	0.003
0.01 <MAF≤ 0.025	0.849	0.140	0.011
0.025< MAF≤ 0.05	0.801	0.174	0.025
0.05 <MAF≤ 0.1	0.725	0.218	0.057
0.1 <MAF≤ 0.5	0.438	0.300	0.262
0.005 <MAF≤ 0.05	0.749	0.234	0.017
0.05 <MAF≤ 0.5	0.477	0.292	0.232

[1] Proportions were calculated for 100 replicates of simulated haplotypes one megabase in length using CEU haplotype reference panel to impute SNPs in the OOA
P(AA|Aa) = the probability that a truly heterozygous genotype was imputed incorrectly as a major allele homozygous genotype
P(Aa|AA) = the probability that a truly major allele homozygous genotype was imputed incorrectly as a heterozygous genotype
OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency in the reference population

Table 3.5. Mean imputation accuracy for chromosome 22 data

Reference population[1]	CEU		CEU	
Study population[2]	OOA		CEU	
	r^2	HetAgree	r^2	HetAgree
0.01 <MAF≤ 0.025	0.47	0.36	0.64	0.26
0.025 <MAF≤ 0.05	0.51	0.34	0.69	0.41
0.05 <MAF≤ 0.1	0.58	0.36	0.74	0.41
0.01 <MAF≤ 0.05	0.50	0.35	0.67	0.36

[1] Ancestry of the haplotype reference panel
[2] Ancestry of the individuals for which genotypes are imputed
OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency in the reference population
r^2 = Squared correlation between the true genotype and the most likely imputed genotype
HetAgree = Proportion of truly heterozygous genotypes that were imputed correctly
The values reported are the means of measures of imputation accuracy for 200 independent imputation analyses of chromosome 22 for each reference/study configuration. For each minor allele frequency bin, the minimum number of SNPs was 1,000.

Table 3.6. Inter-Quartile range for r^2 measure of imputation accuracy for chromosome 22 data

Reference population[1]	CEU		CEU	
Study population[2]	OOA		CEU	
	1stQ	3rdQ	1stQ	3rdQ
0.01 <MAF≤ 0.025	0.08	0.81	0.27	0.95
0.025 <MAF≤ 0.05	0.15	0.86	0.45	0.99
0.05 <MAF≤ 0.1	0.25	0.90	0.56	0.98
0.01 <MAF≤ 0.05	0.12	0.85	0.38	0.98

[1] Ancestry of the haplotype reference panel
[2] Ancestry of the individuals for which genotypes are imputed
OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency in the reference population
r^2 = Squared correlation between the true genotype and the most likely imputed genotype
The values reported are the first quartile and third quartile of the distribution of the r^2 measure of imputation accuracy for chromosome 22 for each reference/study configuration. For each minor allele frequency bin, the minimum number of SNPs was 1,000.

Table 3.7. Comparison of minor allele frequency in simulated haplotypes

	CEU				
OOA	MAF=0	0<MAF<0.5%	0.5≤MAF<1%	1≤MAF<5%	MAF≥5%
MAF=0	-	26.7	0.6	0	0
0<MAF<0.5%	17.4	12.3	1.1	0.1	0
0.5≤MAF<1%	1.7	1.8	3.7	0.5	0
1≤MAF<5%	0.1	0.2	0.7	10.0	0.2
MAF ≥5%	0	0	0	0.2	22.7

OOA = Old Order Amish
CEU = U.S. Utah residents from HapMap
MAF = Minor Allele Frequency in the reference population
Cross classification of the percent of variable sites for different minor allele frequency bins for 981,159 variable sites (in either the CEU or the OOA) in 100 realizations simulating 800 CEU-like and 800 OOA-like haplotypes 1 megabase in length.

3.9 References

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB and others. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52-8.

Beiler K, editor. 1988. Fisher Family History. Third ed: Eby's Quality Printing, PA.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415-25.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-73.

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. 2009. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84(2):235-50.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2):337-8.

Kingman JFC. 1982. On the Genealogy of Large Populations. Journal of Applied Probability 19A:27-43.

Lee WJ, Pollin TI, O'Connell JR, Agarwala R, Schaffer AA. 2010. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. BMC Med Genet 11:68.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu Rev Genomics Hum Genet 10:387-406.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34(8):816-34.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-53.

Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, Jaquish C, Douglas JA, Roy-Gagnon MH, Sack P and others. 2008. The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am Heart J 155(5):823-8.

Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, Horenstein RB, Post W, McLenithan JC, Bielak LF, Peyser PA and others. 2008. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 322(5908):1702-5.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576-83.

Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L, Salomaa V, Daly M, Palotie A, Peltonen L and others. 2010. Founder population-specific

HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res 20(10):1344-51.
Thompson EE, Sun Y, Nicolae D, Ober C. 2010. Shades of gray: a comparison of linkage disequilibrium between Hutterites and Europeans. Genet Epidemiol 34(2):133-9.
Van Hout CV, Levin AM, Rampersaud E, Shen H, O'Connell JR, Mitchell BD, Shuldiner AR, Douglas JA. 2010. Extent and distribution of linkage disequilibrium in the Old Order Amish. Genet Epidemiol 34(2):146-50.

Chapter 4. Genomic estimates of inbreeding in the Old Order Amish

4.1 Abstract

Using a hidden Markov model and genome-wide map of Single Nucleotide Polymorphisms (SNPs), I estimated inbreeding for 837 Old Order Amish individuals who were recruited in the course of the HAPI Heart study [Mitchell, et al. 2008]. I observed strong agreement between genomic and pedigree-based estimates. Moreover, both measures indicated that the mean inbreeding coefficient for these study participants is approximately 0.035, or similar to the expected inbreeding coefficient for the offspring of half 1st cousins. Using SNP specific probabilities of autozygosity, I further characterized the number and lengths of autozygous segments in the study participants. Additionally, I demonstrated that the preponderance of autozygosity in the OOA is likely due to many distant inbreeding loops, consistent with the OOA tradition of avoiding marriage between close relatives.

4.2 Introduction

The Old Order Amish community of Lancaster County Pennsylvania (OOA) is a population founded by approximately 200 individuals of Western European ancestry in the early 1700s [Cross 1964]. Over approximately 15 generations, the OOA have grown to a census size of approximately 35,000 individuals [Agarwala, et al. 2001; Lee, et al.

2010]. Their pedigree has been meticulously documented in the Fisher Book [Beiler 1988], and subsequently incorporated into the Anabaptist Genealogy Database (AGDB) [Agarwala, et al. 2001].

The population demographic history of the OOA lends itself to studies of rare recessive traits [McKusick, et al. 1964]. Since the mid 1960s, the OOA have been the the population of inference for gene mapping studies of rare metabolic disorders [Strauss and Puffenberger 2009]. Despite the fact that detection and characterization of long homozygous and autozygous regions in humans has received considerable theoretical and empirical attention for both cosmopolitan populations [Broman and Weber 1999; Clark 1999; Gibson, et al. 2006; Li, et al. 2006] and population isolates [Chapman and Thompson 2002; Leutenegger, et al. 2003; Sheffield, et al. 1998; Wang, et al. 2006], no genome-wide characterization of inbreeding and the distribution of autozygous segments in the OOA has been published.

In the OOA, inbreeding is driven by demographic history, namely, population growth from a moderate number of founders and limited exogamy. The distinctive cultural identity of the OOA, including non-proselytizing religious traditions and marriage within the church, are considerable barriers to gene flow from more cosmopolitan populations. Practically all of the Old Order Amish are related, [Lee, et al. 2010] meaning that any two individuals inherited a proportion of their genomes from a recent common ancestor. Specifically, the expected proportion of the genome that is shared identical by descent (IBD) between two individuals is defined by their kinship coefficient [Lange 1997; Malécot 1948]. By extension, offspring of related individuals can inherit loci for which both alleles are descended from a single recent ancestor, i.e.,

IBD, and the individual is inbred. The inbreeding coefficient is defined as the probability that two alleles at a locus in an individual are IBD. However, cryptic relatedness, due to errors or omissions in the pedigree, or by undocumented relatedness between the founders of the OOA, may result in underestimated inbreeding.

Accurate knowledge of inbreeding is important in the design, execution and interpretation of genetic epidemiological studies in isolated populations, such as the OOA. For example, underestimation of inbreeding can produce false positive results in the context of linkage analyses [Miano, et al. 2000]. Also, in models of the architecture of complex traits, inbreeding results in a redistribution of variance components, specifically, decreasing narrow sense heritability [Falconer 1989]. Additionally, inbreeding inflates estimates of linkage disequilibrium (LD) [Zhang, et al. 2004] and increases homozygosity compared to Hardy Weinberg Expectation (HWE) [Song and Elston 2003].

To address the possibility of underestimation of inbreeding due to incomplete and/or inaccurate pedigree information in the OOA, I estimated inbreeding coefficients from only genomic data, i.e., independent of any pedigree information. Furthermore, I describe the distributions of counts and lengths of IBD, and provide evidence that the vast majority of autozygosity in the OOA is likely due to multiple inbreeding loops that are deep in the pedigree, i.e., more distant than offspring of 2nd cousins.

4.3 Study participants and genotype data

OOA study participants (n=868) were initially recruited to participate in the Heredity and Phenotype Intervention (HAPI) Heart Study [Mitchell, et al. 2008], which

was designed to detect genetic loci that interact with environmental exposures to modify risk factors for cardiovascular disease. Recruitment efforts focused on multigenerational families of relatively healthy adult men and women. Genome-wide Single Nucleotide Polymorphisms (SNPs) for 861 participants were measured using the Affymetrix 500k genotyping array. Quality control measures are described in detail elsewhere [Van Hout, et al. 2010]. Briefly, SNPs with >5% missing data, >5 Mendelian errors, more than one duplicate inconsistency based on the 61 OOA individuals who were also genotyped on the Affymetrix 6.0 array, or deviation from Hardy Weinberg Expectation (HWE) with p-value $< 10^{-6}$ were omitted. After Quality Control (QC) measures, >85% of autosomal SNPs (415,440 of 489,922) on the chip were retained for downstream analysis. Additionally, study participants with more than 5% missing data (n=9), males with >1% heterozygous SNPs on the X chromosome (n=5), and individuals with relationship discrepancies (n=16) were identified. By failing to meet one or more of these criteria, 24 individuals were omitted, with a final count of 837 study participants for further analysis.

4.4 Methods and statistical analyses

Inbreeding coefficients conditional on the pedigree, F_{Ped}, for 837 study participants were computed from pedigree information using PedHunter version 2.0 [Agarwala, et al. 1998; Lee, et al. 2010] and the AGDB version 5 [Agarwala, et al. 2001] which includes of all known paths of descent between the parents of the study participants.

I estimated the genomic inbreeding coefficient, F_{Geno}, of each study participant using a hidden Markov model and Markov chain Monte Carlo, as implemented in the

program FEstim [Leutenegger, et al. 2003]. Briefly, SNPs for each individual are modeled by a hidden Markov chain with IBD status (autozygous or allozygous) as the hidden state. The model is parameterized in terms of the inbreeding coefficient, *F*, and the rate of change in IBD status per centiMorgan (cM), *A*, where F=0.1 and A=0.2 were used for this analysis. The parameters *F* and *A* and their 95% confidence intervals (CIs) were estimated via maximum likelihood for each individual from the Markov model using Baum's algorithm [Baum L.E. 1970]. In addition to the genome-wide estimates, the posterior probability of autozygosity was estimated for each genotype. This model assumes that population allele frequencies and map positions at each SNP are known and that SNPs are in linkage equilibrium.

For the 415,440 autosomal SNPs that passed QC, allele frequencies and pair-wise estimates of LD between SNPs on the same chromosome were estimated from of a subset of 60 minimally related Amish individuals as previously described [Van Hout, et al. 2010]. Briefly, a subset of 60 individuals were selected from the HAPI study using a simulated annealing algorithm as implemented in the program PedMine [Douglas and Sandefur 2008] to minimize the maximum pair-wise kinship coefficient between individuals in the subset. The maximum kinship coefficient was approximately 0.047 or less than that of 1st cousins, with average kinship 0.029. By comparison, the maximum kinship for a random set of 60 individuals from the HAPI study population was 0.16. The average and maximum inbreeding coefficient (F_{Ped}) for the set of 60 minimally related individuals did not meaningfully differ from those of the complete study (data not shown).

To estimate inbreeding from genomic information, I selected an informative map of low LD SNPs using a windowing approach as implemented in the software PLINK [Purcell, et al. 2007]. For computational efficiency, I maximized the information content by considering only SNPs with MAF > 0.35. SNPs were pruned such that the r^2 measure of pair-wise LD was less than 0.2 for all pairs of SNPs within a window of 200 SNPs. The window was shifted by 20 SNPs, and the pruning process was repeated for the next 200 SNPs. The resulting set of 12,201 SNPs had an average minor allele frequency of 0.44 and average inter-SNP distance of 0.29 cM. The mean pair-wise r^2 between all SNPs on the same chromosome in the map was 0.018. Moreover, 90% of these pairs had r^2 less than 0.051 and 99% had r^2 less than 0.13. Map position information was provided by the array manufacturer [Affymetrix].

Since a detailed assessment of the statistical properties of F_{Geno} as estimated by FEstim has not been published, I implemented a gene-dropping strategy to evaluate the estimator F_{Geno}. Genotypes and founder labels for 12,000 SNPs (MAF 0.45) in linkage equilibrium were simulated for the offspring of 1st cousins according to Mendel's laws, as implemented in the genedrop program of MORGAN [Thompson 2005]. The true inbreeding coefficient, F_{True}, was defined as the proportion of genotypes where both alleles were derived from the same founder. Genomic estimates of inbreeding, F_{Geno}, using only the 12,000 SNPs were estimated using the FEstim algorithm. I simulated genotype data for 1000 independent replicates.

I also used the simulated genotypes and founder allele labels to examine the extent to which the hidden state of the Markov chain (autozygous or allozygous) is captured by the posterior probability (PP) of autozygosity at each SNP across a range of

PP thresholds. For each of 12,000 SNPs in 1000 replicates of the offspring of 1st cousins, the sensitivity (probability that PP of autozygosity ≥ threshold given the SNP was truly autozygous) and specificity (probability that PP of autozygosity < threshold given the SNP was truly allozygous) of the FEstim method to detect autozygosity were estimated for a range of posterior probability thresholds.

To characterize the number and length of autozygous segments from SNPs, I used the PP of autozygosity at each SNP as estimated by FEstim to infer multi-SNP segments of the genome where two alleles are likely to be IBD, i.e., autozygous. Segments were inferred as autozygous if two or more sequential SNPs on the same chromosome had a PP of autozygosity ≥ 0.7.

I used pedigree information to determine whether any of the OOA study participants were the offspring of closely related individuals, specifically, the offspring of 1st cousin or 2nd cousin matings. For each study participant, I compared the number of great-grandparents and great-great-grandparents of a non-inbred individual, 8 and 16, respectively, to the number of unique ancestors in the OOA pedigree. For example, individuals who are the offspring of 1st cousins have exactly six unique great-grandparents, in contrast to eight for a non-inbred individual.

4.5 Results

4.5.1 Genomic and pedigree-based estimates of inbreeding

The mean [range] inbreeding coefficient conditional on the OOA pedigree (F_{Ped}) for 837 OOA study participants was 0.034 [0.0003-0.076] [Table 4.1], which is

approximately equivalent to the expected inbreeding coefficient for the offspring of half 1st cousins.

In the analysis of genomic estimates of inbreeding, the FEstim algorithm converged for approximately 99% of study participants (835 of 837). The pedigree derived estimates of inbreeding for the two study participants whose genomic estimates failed to converge were less than 0.0003, representing the two lowest inbreeding coefficients of all participants in the study. These two individuals did not converge for different initial values of *F* and *A* were omitted from further analyses. Point estimates of inbreeding for the other 835 study participants did not meaningfully change for different initial values of *F* and *A* (data not shown).

The mean [range] of F_{Geno} was 0.036 [0.003 – 0.102] [Table 4.1]. Comparing F_{Geno} and F_{Ped}, the mean [range] within individual difference was approximately 0.002 [-0.048 – 0.044]. For approximately 89% of the study participants (745 of 835), the 95% confidence interval for F_{Geno} contained the expected inbreeding coefficients conditional on the pedigree. Furthermore, for approximately 94% of the study participants (788 of 835) F_{Geno} was significantly greater than zero at α=0.05.

4.5.2 Evaluation of FEstim algorithm

I evaluated the performance of the FEstim algorithm from the simulated genotypes and founder allele labels that were generated via gene dropping on the offspring of 1st cousins pedigree. The mean within individual difference between F_{Geno} and F_{True} was 0.0015 [Table 4.2], indicating a small positive bias in the genomic estimate of inbreeding. I compared the statistical efficiency, i.e., the ability of the estimator to

capture truth, between the genomic estimator and the pedigree-based expectation, i.e., given that the expected inbreeding coefficient (F_{Ped}) for the offspring of 1st cousins is 0.0625. The mean absolute difference (MAD), i.e., $|F_{Geno} - F_{True}|$, was 0.002, while the MAD of between F_{Geno} and F_{Ped} was 0.019. For approximately 92% of replicates, ($F_{Geno} - F_{True}$) was less than ($F_{Geno} - F_{Ped}$). Furthermore, the 95% confidence intervals for F_{Geno} contained F_{True} in 100% of replicates, while only approximately 90% (897 of 1000) contained the expected inbreeding coefficient F_{Ped}.

4.5.3 Inference of autozygous segments

I inferred autozygous segments for the 835 OOA study participants using a PP threshold of autozygosity of at least 0.7 for at least two consecutive SNPs. At this threshold, sensitivity and specificity were 96.0% and 99.9%, respectively, for the simulated offspring of 1st cousins [Table 4.3]. Summary statistics for the number and lengths of IBD segments inferred in the OOA study participants are provided in Table 4.3. The mean total length of inferred IBD segments was approximately 107 cM, or approximately 3.1% of the length of a 35 Morgan genome. The maximum total length IBD for an individual in the study was approximately 328 cM, or approximately 9.4% of a genome of length 35 Morgans. Despite both the negative bias of 96% sensitivity in the identification of truly autozygous SNPs and the small positive bias in the F_{Geno} estimator, these proportions were similar to the mean and maximum estimates of F_{Geno} of 0.034 and 0.102 [Table 4.1].

I used the OOA pedigree to determine whether study participants were the offspring of 1st cousins and/or 2nd cousins by comparing the observed number of unique

ancestors to the expected number of ancestors in a non-inbred pedigree. I identified 152 OOA individuals for whom the most recent inbreeding loop was the offspring of 2nd cousins. For completeness, zero offspring of 1st cousins, 2 offspring of half 2nd cousins, 15 offspring of double 2nd cousins and one offspring of triple 2nd cousins were also identified from the pedigree information. The impact of 2nd cousin matings on inbreeding in the OOA can be characterized by subtracting the expected inbreeding due to known 2nd cousin matings, if any, from each study participant's inbreeding coefficient conditional on the entire pedigree. The mean F_{Ped} is 0.034 [Table 4.1], whereas the mean F_{Ped} after removing the contribution of all 2nd cousin matings is 0.031.

4.6 Discussion

This report is the first comparison of genomic and pedigree-based inbreeding in the OOA. Strong agreement between genomic and pedigree based inbreeding implies that undocumented relationships and inaccuracies in the pedigree have had a small aggregate effect on the inbreeding coefficients of the OOA. Additionally, mean inbreeding coefficients as estimated by both methods are approximately 0.035, or approximately equivalent to the offspring of half 1st cousins (0.031).

There were no observed offspring of 1st cousin matings, and for approximately 80% of (667 of 837) study participants, pedigree information indicates that the most recent inbreeding loop was more distant than the offspring of 2nd cousins. By subtracting the inbreeding coefficient due to offspring of 2nd cousin loops from F_{Ped}, I observed only a small change in the average inbreeding coefficient in all study participants, from 0.034 to 0.031. This demonstrates that the aggregate contribution of many distant inbreeding

loops, i.e., offspring of 3rd cousin and more distant, constitutes approximately 92% (0.031 / 0.034) of the autozygosity in these study participants. This observation is consistent with the OOA practice of avoiding 1st cousin marriages [McKusick, 1978].

The proportion of the study participants who were inferred to be autozygous was calculated for each SNP. At a PP threshold of 0.7, SNPs were most likely to be autozygous in approximately 3.3% of the study population, consistent with the mean genome-wide estimate of inbreeding of approximately 3.5%. Notably, the minimum number of individuals for which any of the 12,201 SNPs was inferred as autozygous was 8 [data not shown]. Single SNPs are very likely to be embedded in autozygous segments that are many cM in length. For example, the mean length of an inferred segment with two alleles IBD in the OOA study participants was 9.7 cM [Table 4.3] and includes on average 33 SNPs given an average inter-SNP distance of approximately 0.29 cM. Though it is possible that small regions of the genome that do not tolerate autozygosity have escaped detection, no such regions were observed in the course of this analysis. Regions of the genome that do not tolerate autozygosity may, for example, be incompatible with life, and would be interesting candidates for further study. The lack of evidence for such regions in the HAPI study sample is noteworthy insofar as there are very few populations with genetic data and suitable demographic histories in which a similar analysis would be possible.

Inference of IBD in the OOA is limited, in part, by a balance between fulfilling the assumption of the FEstim method that SNPs are in linkage equilibrium and the density of SNPs available for analysis. While relatively small violations of the assumption of linkage equilibrium between SNPs inflate estimates of inbreeding

[Polasek, et al. 2010], it is impractical to completely eliminate LD in a map of SNPs that are sufficiently dense for the genomic estimation of inbreeding in the OOA. However, despite sources of error in the genomic estimator of inbreeding, including residual LD between the SNPs and the small positive systematic bias (described in results), simulations indicate that the F_{Geno} estimator captures the true inbreeding substantially more accurately than F_{Ped}.

The limited genetic map resolution also impacts the lower limit of detection of the length of autozygous segments. For example, the expected inbreeding coefficient for the offspring of 5^{th} cousins is 0.5^{12}, which, on average, results in a total autozygous length of approximately 0.8 cM, given a sex averaged map of about 35 Morgans, and is likely to be inherited as a single autozygous region. Autozygous segments of this size are unlikely to be detected using the methods described in this analysis because as the length of a segment, and correspondingly, the number of contiguous homozygous SNPs decreases, the likelihood of homozygosity due to autozygosity approaches the likelihood of homozygosity due to random chance.

4.7 Conclusions

For 835 Old Order Amish individuals, I estimated inbreeding coefficients from genomic information, and observed strong agreement between genomic and pedigree-derived estimates. The mean inbreeding coefficient using genome-wide SNPs and pedigree information was approximately 0.035 by both approaches, and approximately equivalent to the offspring of half 1^{st} cousins. I have also summarized the number and lengths of autozygous segments in the OOA study population. Based on pedigree

information, I show that there are no inbreeding loops closer than that of the offspring of 2^{nd} cousins. Additionally, the contribution of all offspring of 2^{nd} cousin inbreeding loops is approximately 8% of the mean inbreeding in the OOA, indicating that the majority of inbreeding in this study population is due to the aggregate effect of many distant loops, specifically, more distant than the offspring of 2^{nd} cousins.

4.8 Acknowledgements

I thank Drs. Braxton Mitchell, Alan Shuldiner, and Toni Pollin at the University of Maryland for their collaboration on the HAPI Heart Study, and Drs. Alejandro Schaffer, Richa Agarwala and Woei-Jyh (Adam) Lee at the NIH/NCBI for providing the pedigree information and supplying inbreeding coefficients conditional on the entire AGDB. I gratefully acknowledge the Amish Research Clinic Staff, our Amish liaisons, and the Amish community, whose support and cooperation made this study possible. This research was supported by NIH grants T32 HG00040, U01 HL0201, and R01 (CA122844).

4.9 Tables

Table 4.1. Distribution of genomic and pedigree-based estimates of inbreeding in the OOA

	F_{Geno}	F_{Ped}	$F_{Geno} - F_{Ped}$
Minimum	0.003	0.0003	-0.0475
1st Quartile	0.024	0.027	-0.0110
Mean	0.036	0.034	-0.0018
Median	0.034	0.034	-0.0004
3rd Quartile	0.024	0.041	0.0078
Maximum	0.102	0.076	0.0443
MAD			0.0115

The genomic estimate of inbreeding (F_{Geno}) was estimated for 835 OOA study participants using only SNP data from the study participants, i.e., ignoring the pedigree. The pedigree inbreeding coefficient (F_{Ped}) is an expectation conditional on all known paths of descent between the parents of HAPI study participants from the AGDB.
OOA = Old Order Amish
MAD = Mean Absolute Difference, i.e., $| F_{Geno} - F_{Ped} |$.

Table 4.2. Distribution of genomic-derived and actual inbreeding coefficients for simulated offspring of 1st cousins

	F_{Geno}	F_{True}	$F_{Geno} - F_{True}$	$F_{Geno} - F_{Ped}$
Minimum	0.016	0.008	-0.004	-0.047
1st Quartile	0.049	0.047	0.000	-0.014
Mean	0.065	0.064	0.002	0.001
Median	0.063	0.061	0.001	0.001
3rd Quartile	0.080	0.079	0.003	0.018
Maximum	0.172	0.170	0.011	0.110
MAD			0.002	0.019

Genotypes (n=12,000) in linkage equilibrium were simulated for the offspring of 1st cousins via gene dropping. The true inbreeding coefficient (F_{True}) is defined as the proportion of genotypes in an individual with two identical founder labels. The genomic estimate of inbreeding (F_{Geno}) was estimated by the FEstim algorithm using only simulated SNPs. The results of 1,000 replicates are shown.
F_{Ped} = Inbreeding conditional on pedigree for offspring of 1st cousins
MAD = Mean Absolute Difference, i.e., $| F_{Geno} - F_{True} |$ and $| F_{Geno} - F_{Ped} |$

Table 4.3. Classification of marker autozygosity by posterior probability thresholding

Threshold	Sensitivity[1]	Specificity[2]
0.1	98.9	99.2
0.2	98.6	99.5
0.3	98.4	99.6
0.4	98.2	99.7
0.5	97.3	99.7
0.6	96.4	99.8
0.7	96.0	99.9
0.8	93.8	99.9
0.9	91.1	99.96

To assess the sensitivity and specificity of autozygosity for a range of posterior probability (PP) thresholds, founder allele labels and genotypes were simulated for 12,000 SNPs (MAF 0.45) in linkage equilibrium for 1000 offspring of 1st cousin matings via gene dropping. True autozygosity was determined by comparing the founder allele labels at each SNP. The posterior probability of autozygosity for each SNP was estimated using a hidden Markov model as implemented in FEstim.

[1] Sensitivity is defined as the percent of markers for which the PP of autozygosity was ≥ PP threshold given that the marker was truly autozygous.

[2] Specificity is defined as the percent of markers for which the PP of autozygosity was < PP threshold given that the marker was truly allozygous.

Table 4.4. Distribution of segments IBD in OOA study participants

	Mean	Range
Number of segments IBD per individual	10.5	[1 – 26]
Length per IBD segment (cM)	9.7	[3.6 – 25.8]
Longest IBD segment per individual (cM)	24.1	[4.3 – 92.4]
Total length IBD per individual (cM)	107.2	[5.1 – 322.8]

For n=835 study participants, the posterior probability of autozygosity for 12,201 SNPs were estimated in a hidden Markov Model. Autozygous segments were inferred from the posterior probability of autozygosity of at least 0.7 for at least two consecutive SNPs on the same chromosome.

OOA = Old Order Amish

4.10 References

Affymetrix. http://www.affymetrix.com/support/technical/byproduct.affx?product=500k.

Agarwala R, Biesecker LG, Hopkins KA, Francomano CA, Schaffer AA. 1998. Software for constructing and verifying pedigrees within large genealogies and an application to the Old Order Amish of Lancaster County. Genome Res 8(3):211-21.

Agarwala R, Schaffer AA, Tomlin JF. 2001. Towards a complete North American Anabaptist Genealogy II: analysis of inbreeding. Hum Biol 73(4):533-45.

Baum L.E. PT, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematics and Statistics 41(1):164-171.

Beiler K, editor. 1988. Fisher Family History. Third ed: Eby's Quality Printing, PA.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63(3):861-9.

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

Chapman NH, Thompson EA. 2002. The effect of population history on the lengths of ancestral chromosome segments. Genetics 162(1):449-58.

Clark AG. 1999. The size distribution of homozygous segments in the human genome. Am J Hum Genet 65(6):1489-92.

Cross H. 1964. Population studies and the Old Order Amish. Nature 262(5563):17-20.

Douglas JA, Sandefur CI. 2008. PedMine--a simulated annealing algorithm to identify maximally unrelated individuals in population isolates. Bioinformatics 24(8):1106-8.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. Am J Hum Genet 67(5):1219-31.

Falconer DS, MacKay T. 1989. Introduction to quantitative genetics. Burnt Mill, Harlow, Essex, England, New York: Longman, Wiley.

Gibson J, Morton NE, Collins A. 2006. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15(5):789-95.

Lange K. 1997. Mathematical and statistical methods for genetic analysis. New York: Springer.

Lee WJ, Pollin TI, O'Connell JR, Agarwala R, Schaffer AA. 2010. PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. BMC Med Genet 11:68.

Leutenegger AL, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73(3):516-23.

Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MT and others. 2006. Long contiguous stretches of homozygosity in the human genome. Hum Mutat 27(11):1115-21.

Malécot G. 1948. Les mathématiques de l'hérédité. Paris: Masson.

McKusick VA, Hostetler JA, Egeland JA. 1964. Genetic Studies of the Amish, Background and Potentialities. Bull Johns Hopkins Hosp 115:203-22.

McKusick, VA. Medical Genetic Studies of the Amish, Selected Papers. Baltimore; Johns Hopkins Press. 1978.

Miano MG, Jacobson SG, Carothers A, Hanson I, Teague P, Lovell J, Cideciyan AV, Haider N, Stone EM, Sheffield VC and others. 2000. Pitfalls in homozygosity mapping. Am J Hum Genet 67(5):1348-51.

Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, Jaquish C, Douglas JA, Roy-Gagnon MH, Sack P and others. 2008. The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am Heart J 155(5):823-8.

Polasek O, Hayward C, Bellenguez C, Vitart V, Kolcic I, McQuillan R, Saftic V, Gyllensten U, Wilson JF, Rudan I and others. 2010. Comparative assessment of methods for estimating individual genome-wide homozygosity-by-descent from human genomic data. BMC Genomics 11:139.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-75.

Sheffield VC, Stone EM, Carmi R. 1998. Use of isolated inbred human populations for identification of disease genes. Trends Genet 14(10):391-6.

Song K, Elston RC. 2003. Tests for a disease-susceptibility locus allowing for an inbreeding coefficient (F). Genetica 119(3):269-81.

Strauss KA, Puffenberger EG. 2009. Genetics, medicine, and the Plain people. Annu Rev Genomics Hum Genet 10:513-36.

Thompson EA. 2005. MCMC in the analysis of genetic data on pedigrees in Markov chain Monte Carlo: innovations and applications. In: Liang F. WJS, Kendall, W., editor. Lecture note series of the IMS, National University of Singapore. Singapore: World Scientific. p 183-216.

Van Hout CV, Levin AM, Rampersaud E, Shen H, O'Connell JR, Mitchell BD, Shuldiner AR, Douglas JA. 2010. Extent and distribution of linkage disequilibrium in the Old Order Amish. Genet Epidemiol 34(2):146-50.

Wang H, Lin CH, Service S, Chen Y, Freimer N, Sabatti C. 2006. Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density. Hum Hered 62(4):175-89.

Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE. 2004. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proc Natl Acad Sci U S A 101(52):18075-80.

Chapter 5. Discussion

5.1 Implications of findings

Other studies of population isolates, like the OOA, have observed strong similarity in allele frequency patterns and linkage disequilibrium (LD) profiles in comparison to the HapMap CEU, including the Hutterites [Thompson, et al. 2010] and the isolates of the Dalmatian islands of Croatia [Navarro, et al. 2010]. Additionally, these reports support a fundamental element of the common-disease common-variant hypothesis, namely that common SNPs that contribute to variation in disease phenotypes or quantitative traits in these populations exist in more cosmopolitan populations. A potential downside to this strong similarity, however, is the growing evidence that the predicted increase in power of association studies due to extensive LD in isolated populations [Bonnen, et al. 2006] may not have materialized [Huyghe, et al. 2010], at least with regard to common SNPs.

While the emerging story for common SNPs may not be surprising, the extent to which rare SNPs are shared between the OOA and the CEU may be less intuitive. Currently, the 1000 Genomes Project is measuring rare variation in large enough number of individuals to meaningfully study rare SNPs in a moderate number of populations, with stated goals (1000genomes.org) of cataloging most genetic variation above approximately MAF 0.01 and providing haplotype reference panels both for the purposes

of imputing genotypes in various study populations and combining studies with genotype data from different platforms for meta-analysis. Prior to the investigation described in Chapter 3, it was unclear as to whether the creation of a deeply sequenced haplotype reference panel in the OOA would result in improved imputation as compared to the 1000 Genomes Project resources. The results of this study suggest that, at least using the imputation strategy as implemented in Chapter 3, the creation of a population-specific haplotype reference panel in the OOA is likely to result in only a small gain in imputation accuracy. This information is likely to be useful when contemplating study strategies for measuring rare variants in other isolated populations. More broadly, simulating haplotypes and conducting mock analyses prior to proposing and/or executing genetic sequencing studies may offer important insights and identify potential pitfalls of newer approaches such as imputation.

In chapter 4, I compared estimates of inbreeding derived from genomic data, namely, SNPs to those from pedigree data. I also evaluated the genomic estimator of inbreeding using simulated data, and observed that it captured the true inbreeding coefficient better than the pedigree data. Additionally, both the genomic and pedigree-derived estimates had high agreement, suggesting that even though they may be less precise, the pedigree based estimates capture most of the inbreeding in the OOA, i.e. the pedigree based estimates are not substantially underestimating the true inbreeding in this population. Thus, cryptic relatedness among the founders of the OOA and/or errors and omissions in the pedigree information are unlikely to result in an underestimation of inbreeding.

5.2 Future directions

The comparison of allele frequencies and LD profiles for common SNPs between the OOA and CEU as described in Chapter 2 has proved to be an important resource for applied human genetic studies in the OOA. Though the work in this thesis has focused on applied projects, often in the context of gene mapping, there are theoretical treatments of LD in populations. In particular, models of LD and population size [Sved 2009], and of LD and inbreeding [Haldane 1949; Hill 1975; Sved 1971] have received considerable theoretical attention. It is humbling to consider the anachronism of how early theorists might have dealt with the richness and availability modern genome-wide data.

In the investigation of the accuracy of imputation of rare variants described in Chapter 3, the imputation strategy that was implemented makes no use of the extensive pedigree that exists for the OOA. It is likely that strategies that use prior information about the relatedness of the study population could substantially increase imputation accuracy for rare variants. In fact, there are a number of algorithms and computational strategies that have been developed to incorporate family information for imputation of genotypes [Burdick, et al. 2006; Kirkpatrick, et al. 2010; Meuwissen and Goddard 2010]. However, implementation of these strategies using deep pedigrees like those of the OOA is challenging, though, pedigree trimming approaches [Liu, et al. 2008] which retain many of the 1st and 2nd degree relationships in larger pedigrees, could be useful tools with which to manage the computation complexity of imputation in large families.

Another strategy to potentially increase imputation accuracy might be to maximize the genetic diversity of the haplotype reference panel, e.g., manipulate the composition of the haplotype reference panel to include a set of minimally related

individuals. For example, the simulated annealing algorithm [Douglas and Sandefur 2008] that was implemented to select the set of 60 minimally related individuals for the purposes of allele frequency and LD estimation is an obvious first choice for the identification of unrelated individuals in this context.

Each of the strategies outlined above should be compared to develop a clear idea as to which imputation strategy might have the highest accuracy. Populations with well characterized pedigrees are uniquely suited to accurately identify rare variation because related individuals can be used to reduce uncertainty in measuring rare genotypes, so it is certain that efforts to evaluate the contribution of rare variation to disease phenotypes and quantitative traits will continue to be pursued in populations with well documented genealogies, like the OOA. It is worth noting that until accurate whole genome sequencing becomes practical for large studies, efforts to refine imputation strategies may result in substantial cost savings compared to study strategies that conduct deep sequencing all of the study participants.

In Chapter 4, as part of my exploratory analysis, none of the 12,201 SNPs were inferred (via the PP of autozygosity > 0.7 for at least two SNPs on the same chromosome) to be allozygous for all 835 OOA study participants. Moreover, in this preliminary analysis, the distribution of the proportion of the study participants who were autozygous at a given locus appeared to be consistent with random chance. It is possible that a similar analysis in a larger number of study participants may reveal regions of the genome which are autozygous less often than by chance, possibly due to the lethal recessive action of genes in the region of autozygosity.

5.3 References

Bonnen PE, Pe'er I, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly MJ, Reich DE and others. 2006. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat Genet 38(2):214-7.

Burdick JT, Chen WM, Abecasis GR, Cheung VG. 2006. In silico method for inferring genotypes in pedigrees. Nat Genet 38(9):1002-4.

Douglas JA, Sandefur CI. 2008. PedMine--a simulated annealing algorithm to identify maximally unrelated individuals in population isolates. Bioinformatics 24(8):1106-8.

Haldane JB. 1949. The association of characters as a result of inbreeding and linkage. Ann Eugen 15(1):15-23.

Hill WG. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor Popul Biol 8(2):117-26.

Huyghe JR, Fransen E, Hannula S, Van Laer L, Van Eyken E, Maki-Torkko E, Lysholm-Bernacchi A, Aikio P, Stephan DA, Sorri M and others. 2010. Genome-wide SNP analysis reveals no gain in power for association studies of common variants in the Finnish Saami. Eur J Hum Genet 18(5):569-74.

Kirkpatrick B, Halperin E, Karp RM. 2010. Haplotype inference in complex pedigrees. J Comput Biol 17(3):269-80.

Liu F, Kirichenko A, Axenovich TI, van Duijn CM, Aulchenko YS. 2008. An approach for cutting large and complex pedigrees for linkage analysis. Eur J Hum Genet 16(7):854-60.

Meuwissen T, Goddard M. 2010. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185(4):1441-9.

Navarro P, Vitart V, Hayward C, Tenesa A, Zgaga L, Juricic D, Polasek O, Hastie ND, Rudan I, Campbell H and others. 2010. Genetic comparison of a Croatian isolate and CEPH European founders. Genet Epidemiol 34(2):140-5.

Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol 2(2):125-41.

Sved JA. 2009. Linkage disequilibrium and its expectation in human populations. Twin Res Hum Genet 12(1):35-43.

Thompson EE, Sun Y, Nicolae D, Ober C. 2010. Shades of gray: a comparison of linkage disequilibrium between Hutterites and Europeans. Genet Epidemiol 34(2):133-9.