University of Calabria

PhD Program in Operation Research

Genetic analysis of neuroblastoma in African-Americans

Candidate
Valeria Latorre

Supervisor
Prof. Giuseppe Passarino

Co-ordinator
Prof. Lucio Grandinetti

INDEX

Introduction

During my PhD program my interest has been addressed to the statistical analysis of two different complex phenotypes: longevity and neuroblastoma.
In the first year I participated in the study "The genetic component of human longevity: analysis of the survival advantage of parents and siblings of Italian nonagenarians" where our aim was to estimate the genetic component of longevity from families of nonagenarians from a population of southern Italy, Calabria. We analyzed the survival functions comparing parents and siblings of long-lived subjects to the appropriate Italian birth cohorts and siblings to their spouses.
The reduced mortality of relatives of centenarians has suggested the presence of a genetic component in the longevity trait. Heritability of a trait is population specific and it may be influenced by different factors acting differently on certain traits in different populations such as living in areas with slower progress (such as Calabria). We conducted an analysis on parents and siblings of the probands, finding that they both have a significant survival advantage over their Italian birth cohort counterparts, but female siblings did not show the advantages that males did.
Our results should be read with some considerations as to the largely rural and underdeveloped society where our data comes from, which has maintained strong social differences until a few decades ago.
In the subsequent and last two years of my PhD course I collaborated on a GWAs project at the Children's Hospital of Philadelphia (CHOP) resulting in a work called "Replication and fine mapping of neuroblastoma SNP

association at the *BARD1* locus in African-Americans". We analyzed data from the African-American population to replicate results from a previous analysis on Caucasian to find genetic association with neuroblastoma.
We performed both a GWAs analysis and a Candidate Gene analysis. Our results confirmed our expectations but only for the BARD1 gene. To increase the power and have more reliable result we need to increase the sample size and to replicate results in other populations.

In the following chapters I will present in details the work on GWAs analysis for mapping neuroblastoma genes in African-Americans. In the appendix the published paper "The genetic component of human longevity: analysis of the survival advantage of parents and siblings of Italian nonagenarians" (European Journal of Human Genetics) is presented.

CHAPTER I
Neuroblastoma: Clinical, epidemiological and molecular aspects

Neuroblastoma is an embryonal cancer arising from any neural crest of the sympathetic nervous system. It is the most common cancer diagnosed during the first year of life and is often lethal.

The cell of origin is thought to be a developing and incompletely committed precursor cell derived from neural-crest tissues.

The neural crest is a group of cells in the early embryo that give rise to many tissues and organs. Cells migrate to form parts of the autonomic nervous system, which controls body functions such as breathing, blood pressure, heart rate, and digestion and also give rise to many tissues in the face and skull, and other tissue and cell types. Although many lower-stage neuroblastomas are encapsulated and can be surgically excised with little chance of complications, higher-stage tumors often infiltrate local organ structures, surround critical nerves and vessels such as the celiac axis, and are largely unresectable at the time of diagnosis. Neuroblastomas typically metastasize to regional lymph nodes and to the bone marrow by means of the hematopoietic system. Tumor cells metastatic to marrow can infiltrate cortical bone. However, transient and complete regression often occurs with no intervention other than supportive care (Westermann et al. 2002, Maris et al. 2007, Henderson et al. 2011).

Neuroblastoma and other cancers occur when a build-up of genetic mutations in critical genes, usually those that control cell proliferation or differentiation, allow cells to grow and divide uncontrollably to form a tumor. In most cases, these genetic changes are acquired during a person's lifetime, and such changes are known as somatic mutations.

Despite the wealth of knowledge about somatically acquired genomic aberrations that correlate with tumor phenotype, little is known about the events that predispose to the development of neuroblastoma.

A common environmental exposure that influences susceptibility to neuroblastoma has been difficult to identify using epidemiologic studies (Maris et al. 2007). Due to the lethality of neuroblastoma in early childhood, genetic studies of hereditary disease have been hampered by the rarity of the condition and the small size of pedigrees (Maris et al. 2008).

The clinical presentation is highly variable, and although a substantial proportion of affected individuals may show a favorable outcome and may even have spontaneous regression of a localized, or even disseminated, tumor. Approximately 50% of cases show an aggressive clinical course with widespread metastatic disease that have survival rates of less than 35% despite aggressive therapy with dose-intensive induction chemotherapy and surgery, followed by myeloablative therapy with stem cell rescue, local radiation therapy and biological response modification using retinoids and/or immunotherapy. Survivors often have serious lifelong coexisting conditions (Maris et al. 2007).

Tumors from patients with an aggressive phenotype resistant to therapy often show focal amplification of the MYCN oncogene or deletions of chromosome arms 1p and 11q, or both. However, because MYCN is so aberrantly dysregulated, and no putative tumor suppressor gene at 1p and 11q has been shown to harbor inactivating mutations in more than a small percentage of cases, no tractable molecular target approaches at present exist for this disease (Mosse et al. 2008).

Until recently little was known about the constitutional genetic events that initiate tumorigenesis, although these somatically acquired genomic alterations are of clinical use as prognostic biomarkers.

On the other hand, tumors showing no structural chromosomal changes but hyperdiploidy due to whole-chromosome gains are more easily cured and

may even spontaneously regress (Wang et al. 2010).

Such different clinical behaviors permit a classification of neuroblastoma into three risk groups: high-risk, intermediate-risk and low-risk (Nguyen 2011).

Patients with widely disseminated disease at diagnosis and a poor survival probability are classified as high-risk, while low-risk patients are those that show favorable clinical features including spontaneous regression of disease with a greater than 95% survival probability with minimal or any chemotherapy. Intermediate-risk cases are the most heterogeneous, and are also the smallest subset using current definitions, comprising about 15% of all neuroblastoma patients.

Neuroblastoma can occur with disorders related to abnormal development of neural-crest-derived tissues such as central congenital hypoventilation syndrome and Hirschsprung's disease. Like most human cancer, neuroblastoma arise sporadically, with less than 1% of cases inherited in an autosomal dominant fashion with a standardized incidence ratio of 9.7 for siblings of index cases. Because of the lethality of the condition before reproductive age, previous genetic linkage scans have been underpowered and results have been difficult to replicate (Mosse et al. 2008).

Rare mutations in the *PHOX2B* gene have been identified in people with neuroblastoma. The *PHOX2B* gene provides instructions for making a protein that acts early in development to help promote the formation of nerve cells and regulate the process by which the neurons mature to carry out specific functions. The protein is active in the neural crest.

Somatic mutations are present only in certain cells and are not inherited and, less commonly, gene mutations that increase the risk of developing cancer can be inherited from a parent. Both types of mutation occur in

neuroblastoma. Somatic mutations in the *PHOX2B* gene increase the risk of developing sporadic neuroblastoma, and inherited mutations in the *PHOX2B* gene increase the risk of developing familial neuroblastoma.

Mutations in the *PHOX2B* gene change a single protein building block in the PHOX2B protein or it could be an addition or deletion of several DNA building blocks in the *PHOX2B* gene. Addition or deletion of nucleotides changes the sequence of amino acids in the PHOX2B protein. All of these types of mutations have been found in familial and sporadic neuroblastoma. The mutations are believed to interfere with the PHOX2B protein's role in promoting nerve cell differentiation, which results in an excess of immature nerve cells and leads to neuroblastoma.

However, PHOX2B mutations explain only a small subset of hereditary neuroblastoma of cases with associated disorders of neural-crest-derived tissues, and are not somatically acquired in tumors, leaving the genetic aetiology for most familial neuroblastoma cases (Mosse et al. 2008).

Germline mutations in the anaplastic lymphoma kinase (ALK) gene explain most hereditary neuroblastomas. It was identified as a familial neuroblastoma gene in 2008 (Mosse et al. 2008) and only half of familial cases that are ~1% of the total.

The fact that no commonly mutated gene has been identified suggest locus heterogeneity with both benign and malignant forms occurring in the same family.

We therefore hypothesize that neuroblastomas arise from relatively common DNA variations that predispose to an increased risk of neuroblastic malignant transformation (Diskin et al. 2009).

References

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K et al. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459, 987-991.

Henderson TO, Bhatia S, Pinto N, London WB, McGrady P, Crotty C et al. (2011) Racial and Ethnic Disparities in Risk and Survival in Children With Neuroblastoma: A Children's Oncology Group Study. *Journal of Clinical Oncology* 29, 76-82

Maris JM, Hogarty MD, Bagatell R, Cohn SL (2007) Neuroblastoma. *Lancet* 369: 2106–2120

Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH, et al. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N Engl J Med* 358, 2585-2593.

Maris JM (2010) Recent Advances in Neuroblastoma. *N Engl J Med* 362, 2202-2211

Mosse YP, Laudenslager M, Longo L, Cole KA, Wood A, Attiyeh EF, et al. (2008) Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* 455, 930- 935

Nguyen LB, Diskin SJ, Capasso M, Wang K, Diamond MA, Glessner J, et al. (2011) Phenotype restricted genome-wide association study using a gene-

centric approach identifies three low-risk neuroblastoma susceptibility loci. *PLoS Genet* 7, e1002026

Wang K, Diskin SJ, Zhang H, Attiyeh EF, Winter C, Hou C, et al. (2011) Integrative genomics identifies LMO1 as a neuroblastoma oncogene. *Nature* 469, 216-220.

Westermann F, Schwab M (2002) Mini-review Genetic parameters of neuroblastomas. *Cancer Letters* 184, 127–147

CHAPTER II
Genetic Association Studies and Genome Wide Association studies (GWAs)

The publication in 2000 of the first draft of the human genome sequence, the deposition of millions of SNPs into public databases, rapid improvements in SNP genotyping technology and the initiation of the International HapMap Project have set the stage for genome-wide association studies, in which a dense set of SNPs across the genome is genotyped to find disease genetic variants for complex traits (Collins et al. 1997).

Genetic variation in DNA sequence affect susceptibility to common diseases and also influence disease-related quantitative traits. Identifying relevant causal genes involved in common 'complex diseases' has proven much more difficult than studies on genetic variations causing phenotypes showing a clear Mendelian segregation. In the latter case the carriers of the mutated alleles carry a quantifiable risk of the disease; on the other hand, in complex phenotypes each causal gene only makes a small contribution to overall heritability, neither necessary nor sufficient to individually cause the disease (Zondervan et al. 2007, Ziegler et al. 2008).

Genetic association studies try to identify correlation between a phenotype, in many cases a binary disease status, and one or more SNPs, genetic markers that are easy to type and that are abundant in the human genome.

Population-based case–control studies and family-based studies are the two main types of genetic association studies (Patterson et al. 2009).

Genetic association study can be Candidate Gene (CG), focusing on a particular gene or area of the genome, or can involve genome-wide association (GWA) analyses conducted without prior hypotheses (Kruglyak 2008, McCarthy et al. 2008). Candidate gene association studies analyze SNPs in candidate genes or regions. These candidates are plausible because of previous study results or on the basis of biological hypotheses, and some of them have strong genetic effects. This approach has sometime been

successful but it is inadequate to fully explain the genetic basis of the disease when its fundamental physiological defects are not known. In addition, many CG studies have not provided evidences for the success or failure of their intended objectives because they have been poorly designed in terms of case definition, control selection, genetic marker selection and particular sample size (Clarke et al. 2011).

More recently, because of reductions in genotyping costs and more sophisticated specifications of the genotyping arrays in terms of SNP numbers and coverage, the potential for GWA studies has been considered. In fact, to fully understand the allelic variation that underlies common diseases, complete genome sequencing for many individuals with and without disease would be required. This is still not technically and economically feasible. However, it has become possible to carry out partial surveys of the genome by genotyping large numbers of common SNPs. In these GWAs, several hundreds of thousands SNPs are analyzed at the same time, posing substantial biostatistical and computational challenges. In fact, as the GWAS approach looks at many SNPs simultaneously and each SNP tested constitutes a separate hypothesis test, so it needs very significant associations and large sample sizes to avoid false positives and to find variants with low odds ratios (Hirschhorn et al. 2005, Frazer et al. 2009, Balding 2006, Wang et al. 2005).

In the design of GWA analysis, as for any case–control study, the first step is to adequately define the disease or phenotype of interest. Nonspecific case definitions could increase both the genetic and the environmental heterogeneity underlying causal factors, decrease the power of detection of an effect and make the replication of the study impossible (Brookfield 2010).

The required size of each study will depend on whether the analysis includes case subgroups; whether the analysis is on candidate genes with a limited number of independent tests or GWAs with many thousands of tests; and whether there is an a priori hypothesis to be tested relating to a polymorphism of known allele frequency.

The variables in a GWA (SNP genotypes) are generated in a highly automated way and this poses further challenges for the quality control of the data. With GWAs hundreds of thousands or even a million genotypes are assessed per individual. In removing false-positive associations, one must undertake several quality control (QC) steps to remove individuals or markers with particularly high error rates. If many thousands of cases and controls have been genotyped to maximize the power to detect association, the removal of a small quantity of individuals, given the large number of markers genotyped in modern GWA studies, should not markedly decrease the overall power of the study (Storey and Tibshirani 2003). The impact of removing one marker is potentially greater than the removal of one individual. It is important to apply quality control per sample and per SNP starting from genotyping errors, especially if occurring differentially between cases and controls. The frequency of missing genotypes for each SNP is another important quality criterion that should be investigated separately for all study groups because of the possibility of differential missingness. The acceptable SNP call rate is typically 95%.

SNPs are often excluded from analysis if the minor allele frequency (MAF) is low. If a SNP has an allele frequency <1% it is reasonable to exclude this SNP because of the low power to detect an association between the SNP and the trait of interest.

The Hardy-Weinberg Equilibrium (HWE) law provides a model that describes and predicts genotype and allele frequencies in a non-evolving population.

Deviations from HWE can be due to inbreeding, population stratification or selection (Balding 2006, Anderson et al. 2010).

Testing for deviations from HWE can be carried out using a Pearson goodness-of-fit test. It is easy to compute, but the $\chi 2$ approximation can be poor when there are low genotype counts, and it is better to use a Fisher exact test, which does not rely on the $\chi 2$ approximation. Typically the test is verified on the control group and SNPs are excluded from further analysis if the p-value is less than 10^(-4).

Odds ratios associated with the risk allele are initially over-estimated, usually in the range of 1.2 to 1.3. It often leads to replication studies that lack sufficient sample size and power to replicate the association because larger samples are needed to detect smaller odds ratios.

Further complexity in the analysis emerges due to the multiple testing carried out in GWA studies. The most common manner of dealing with this problem is to apply the Bonferroni correction, in which the conventional P value is divided by the number of tests performed. This correction assumes independent associations of each SNP with disease even though individual SNPs are known to be correlated due to linkage disequilibrium, so it could result over-conservative (Balding 2006, Dudbridge et al. 2008, Clarke et al. 2011).

Each individual carries 2 copies of each autosomal SNP, so the frequency of each of 3 possible genotypes can be compared in cases and controls, and represented in a contingency table of counts of disease status by either genotype count or allele count (McCarthy et al. 2008).

Under the null hypothesis of no association with the disease, we expect that the relative allele or genotype frequencies is the same in case and control groups. The test of association is a $\chi 2$ test for independence of the rows and columns of the 2×3 contingency table of case-control genotype counts with 2 degrees of freedom, where each of the genotypes is assumed to have an independent association (Clarke et al. 2011).

One important assumption is the presence of HWE in controls. If it is not satisfied, an alternative methods must be used to test for multiplicative models as the Cochran-Armitage trend test (Ziegler et al. 2008).

Tests of association can also be conducted with likelihood ratio (LR) methods where the likelihood of the observed data under the proposed model of disease association is compared with the likelihood of the observed data under the null model of no association.

When there is a need to include additional covariates to handle complex traits, in situations in which the disease risk can be modified by environmental effects such as epidemiological risk factors, population stratification, or interactive and joint effects of other marker loci, more complicated logistic regression models of association are used. In logistic regression models, the logarithm of the odds of disease is the response variable, with linear combinations of the explanatory variables entering into the model as its predictors (Clarke et al. 2011).

Controls should be selected from the same population of the cases, and should be representative of the population who would have become cases according to the case definition and recruitment strategies for the study. Applying this rule spurious findings, due to information and selection biases, and confounding are minimized. It is worth mentioning that in genetic association studies the most important type of bias is related to the ethnic

origin of cases and controls, confounding that is often referred to as population stratification. It could be avoided matching controls to cases on potentially important confounders or adjusting the results for these confounders. Matching is necessary just when the frequency of the confounder shows a big difference between cases and controls. Population stratification, that is, confounding by ethnicity, occurs when the population substructure is not equally distributed between case and control groups. Large sample sizes are required to detect common variants in complex diseases, so a small degree of population stratification can affect a GWA study and association could be found because of allele frequency differences between the founder populations that differentially comprise cases and controls (Ioannidis et al. 2009).

It is useful to remove or reduce the effect of population stratification through the removal of individuals of divergent ancestry. Correction for fine-scale or within-population substructure can be attempted during association testing. The most common method for identifying and subsequently removing individuals with large-scale differences in ancestry is principal component analysis (PCA). An alternative method is multidimensional scaling. For illustration, a Q–Q-plot of all test statistics can be generated showing the degree of inflation of test statistics. Deviations from the diagonal identity line could suggest that the assumed distribution is incorrect and significant differences exist in population structure between cases and controls.

A way of separating the many false-positive associations from the few true-positive associations with disease in GWA studies is the replication of results in independent samples. This is typically included in a single GWA report as part of a multistage design where all SNPs are tested in a random subset of cases and controls, and those exhibiting a nominal predetermined

significance level are taken through to be tested in the remainder of the study sample. Subsequent analysis needs to be carried out for the different stages combined to maintain power level (Altschuler et al. 2008).

When designing a replication study, one should base sample size calculations on a smaller effect size than found in the original study. The two studies should have same or very similar phenotype and population. The replication study will involve analysis of the same SNP and the same allele as the initial one (Goldstein et al. 2009).

Frequently the genetic associations fail to be replicated and this could be attributed to population stratification, phenotype differences, selection biases, genotyping errors, and other factors. The best way of resolving these inconsistencies is to increase the sample sizes, although this may not be feasible for rare conditions or for associations identified in unique populations (Pearson et al. 2008).

GWAS, like all association studies, are largely based on the idea that most alleles are in linkage disequilibrium (LD). Many SNPs have alleles that show strong LD with other nearby SNP alleles so it is possible to perform genome-wide association studies with a selection of SNPs (tag SNPs) that can provide adequate ‘coverage’ of the region in an association study. In order to have information on the LD of different chromosome regions, the Hap Project has been successfully carried out. HapMap provides information on the location of millions of common SNPs across the genome in populations of different ethnic origin and their LD, the allelic association between SNPs located near each other in order to select sets that would efficiently capture untyped common variants to maximize efficiency and power (McCarthy et al. 2008).

LD indicated the non-random association of alleles at nearby genetic loci,

and is the result of shared ancestry where alleles tend to be inherited together on the same chromosome, with specific combinations of alleles known as haplotypes. It permits the analysis of a large number of genes that occur in regions of high LD gaining cost efficiency because it is not necessary to genotype all the SNPs. Choosing a subset of SNPs, or tagSNPs, in strong LD with other SNPs, will capture most of the allelic variation in a region. We suppose that the chosen SNPs may be in close vicinity to functional mutations and therefore associated with the disease because of LD rather than be the cause of the disease. LD is crucial to the design of association studies. If a causal polymorphism is not genotyped, we can still hope to detect its effects through LD with polymorphisms that are typed (Frazer et al. 2009).

To assess the power of a study design we need to measure LD that actually is a non-quantitative phenomenon: there is no natural scale for measuring it. The two most important measures that have been proposed for two-locus haplotype data, are D′ and r2.

For two biallelic loci D is the difference between the observed and expected frequencies of a gamete (haplotype) and D′ is the standardized modification of D. D′ is sensitive to even a few recombinations between the loci and it measures only recombinational history while r2 reflects statistical power to detect association and it summarize both recombinational and mutational history. D′ is inflated with small sample sizes, which is why r2 is often preferred over D′ (Ioannidis et al. 2009, Brookfield et al. 2010).

References

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science*, 322, 881-888.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT (2010) Data quality control in genetic case-control association studies. *Nat Protoc* 5, 1564–1573.

Balding DJ (2006) A tutorial on statistical methods for population association studies *Nature Reviews Genetics* 7, 781-791.

Brookfield JFY (2010) Q&A: promise and pitfalls of genome-wide association studies. *BMC Biology* 8, 41.

Cardon LR (2006) Genetics: Enhanced: Delivering New Disease Genes. *Science* 314, 1403-1405.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G (2007) NCI-NHGRI Working group on replication in association studies, replicating genotype–phenotype associations. *Nature* 447, 655-660.

Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies nature protocols. *Nat Protoc* 6,121-133.

Collins FS, Guyer MS, Chakravarti A (1997) Variations on a Theme: Cataloging Human DNA Sequence Variation. *Science* 278, 1580-1581.

Dudbridge F, Gusnanto A (2008) Estimation of Significance Thresholds for Genomewide Association Scans. *Genet. Epidemiol* 32, 227–234.

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* 10, 241–251.

Goldstein DB (2009) Common Genetic Variation and Human Traits. *N Engl J Med* 23, 1696-1698.

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6, 95-108.

Hongzhe L, Zhi W (2010) A hidden Markov random field model for genome-wide association studies. *Biostatistics* 11, 139–150.

Ioannidis JPA, Thomas G Daly MJ (2009) Validating, augmenting and refining genome-wide association signals *Nature Reviews Genetics* 10, 318-329.

Kruglyak L (2008) The road to genome-wide association studies. *Nature Reviews Genetics* 9, 314-318.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9, 356-369.

Pearson TA, Manolio TA (2008) How to Interpret a Genome-wide Association Study. *JAMA* 299, 1335-1344.

Pettersson FH, Anderson CA, Clarke GM, Barrett JC, Cardon LR, Morris AP, Zondervan KT (2009) Marker selection for genetic case–control association studies. *Nat Protoc* 4, 743-752.

Spix C, Pastore G, Sankila R, Stiller CA, Steliarova-Foucher E (2006) Neuroblastoma incidence and survival in European children (1978–1997): Report from the Automated Childhood Cancer Information System project. *Eur J Cancer* 42, 2081–2091.

Storey JD and Tibshirani R (2003) Statistical significance for genomewide studies. *PNAS* vol. 100, 9440–9445.

Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. *Nature Reviews* 6, 109-118.

Ziegler A, Kænig IR, Thompson JR (2008)Biostatistical Aspects of Genome-Wide Association Studies. *Biometrical Journal* 50, 8-28.

Zondervan KT and Cardon LR (2007) Designing candidate gene and genome-wide case–control association studies. *Nat Protoc* 2, 2492-501.

CHAPTER III
Genetic history of Africans and African Americans

Most of the present day African-Americans are the descendants of the slaves which have been brought to America from Africa through the first half of the XIX century. Although the genetic structure of this population has been influenced by a significant contribution of Caucasians, we need to consider the characteristic of the African population, in order to understand most of the genetic features of African-Americans and, then, to investigate and explain some of the dynamics that led to the genetics of contemporary populations.

African Diversity

The climate in Africa varies from region to region. The different climates range from those of the world's largest desert and second largest tropical rainforest to those of savanna, swamps and mountain highlands, and in the past 10 000 years, in some cases, these climates have gone through dramatic shifts (Reed et al. 2006).

Due to huge environmental diversity, African populations not only differ genetically but they also show a range of linguistic, cultural, and phenotypic variation.

In Africa has been estimated more than 2000 distinct ethno-linguistic groups speaking language that constitute nearly one-third of all languages spoken in the world. These languages have been classified into four major macrofamilies: Niger- Kordofanian, Afroasiatic, Nilo-Saharan and Khoesan. There are few isolated languages which no relationship with all the others has been found (Tishkoff et al. 2009).

Paleontology

As human paleontology says, humans have spread at least twice from Africa to Eurasia. The two migrations involved first archaic populations such as *Homo erectus* and *Homo heidelbergensis* and then anatomically modern humans, even though the transition to modern humans occurred over a substantial period of time and across a broad geographic range within Africa (Reed et al. 2006).

Considering the greater potential for migration and admixture within a single continental region, the hypothesis of a multiregional origin model for modern humans within Africa seems to be likely to have occurred, even though a stronger evidence suggest that East Africa could be the most recent common origin of all modern humans that spread across the rest of the globe within the past ~100,000 years.

Thus, modern humans have existed continuously in Africa longer than in any other geographic region and have maintained relatively large effective population sizes, resulting high levels of genetic variation not only within but also between populations, due to long and short-range of migration events.

Genetic studies results on mtDNA confirm this hypothesis on the historical origin of modern human population (Reed et al 2006).

Although the presence of substructure in the African population can cause spurious results, analysis of African population plays a very important role to characterize the pattern of genetic variation and the relationships among ethnically diverse African populations to reconstruct human evolutionary history (Sankaraman et al. 2008).

Individuals genomes from admixed populations consist of chromosomal segments of distinct ancestry. Estimating individuals local ancestry, number of copies of each ancestry at each location in the genome could have very important applications in disease mapping and in understanding human history.

Meiotic crossover in admixed populations leads to a mosaic of chromosomal segments derived from one or the other ancestral subpopulation. The proportion of admixture and the length of chromosomal segments is influenced by the duration, direction, and rate of gene flow between the two populations, which will vary among individuals.

The gene flow can be a single event in time or continuous over many generations and it results in the temporary generation of long haplotype blocks, which includes polymorphic variants, derived from one or the other ancestral population (Shriner et al. 2010).

In the first few generations of following introgression these blocks of alleles in LD are extremely extended, but with increasing generations they become progressively shorter by recombination. The length of haplotype segments depends on both the number of generations since the initial admixture event and the duration of gene flow (Winkler et al. 2010).

Within the last several hundred years gene flow began between reproductively isolated populations resulting in chromosomal admixture. In admixed populations linked alleles will show extended linkage disequilibrium (LD) relative to the ancestral populations.

Gene flow and resulting admixture have occurred throughout human history, but it is the relatively recent gene flow between continental populations that could bring important results for admixture mapping (Tishkoff et al. 2009, Cheng et al. 2010, Hooker et al. 2010).

African diaspora started at the beginning of 16th century and has resulted in gene flow between previously separated human subpopulations, specifically in two-way admixture between Africans and Europeans in the United States. For a single individidual African ancestry proportions may vary from 100% to 1% African, and on average African Americans have gametes which proportion is approximately 80% African and 20% European (Zheng et al. 2010).

The development of high-throughput SNP genotyping methodologies and the use of elevated levels of LD in recently admixed populations, such as African population, hold great potential for reconstructing patterns of African ancestry among African Americans and for enabling genome-wide association mapping of complex disease susceptibility and pharmacogenomic response in African-American populations (Sankaraman et al. 2008, Yang et al. 2010).

But still little is known about fine- scale population structure at a genome-wide level because previous studies of high-density SNP and haplotype variation among global human populations have included few African populations and in detailed studies of genetic structure among African populations a modest number of markers has been included.

Africa has a complex population history and variation in climate diet and exposure to infection disease. Their populations exhibit a high level of genetic diversity and phenotypic variation that means that there is great potential to find genetic polymorphism at disease susceptibility loci.

Mapping complex disease is more and more getting less expensive and it will facilitate the possibility of doing whole-genome association studies of large number of individuals (Shlush et al. 2010).

Genetics

In studies that estimated admixture proportions in African Americans it was used just a single ancestral African population which is Yoruba. Comparing the inferred West African segments of African-American genomes with contemporary West African populations it was found that the ancestry of the West African component of African Americans is most similar to the profile from non-Bantu Niger-Kordofanian- speaking populations, which include the Igbo, Brong, and Yoruba. All these three populations are likely to have contributed ancestry to present-day African Americans so any of them could be used in admixture mapping studies (Alexander et al. 2009, Bryc et al 2010).

Historical documents show that the Igbo and Yoruba are within the most frequent ethnicities in slave trade records, even though also other African populations not sampled, including those from Sierra Leone, Senegal, Guinea Bissau, and Angola, are also good as ancestral population of some African Americans.

Some self-reported ethnicity as African American show almost no West African ancestry and others show almost complete West African ancestry, these individuals are most likely descendants of individuals of recent African immigrants.

Assuming that the reported ethnicity of these individuals are not mislabeled, it seems that the term African American implies an extremely diverse range of genetic ancestry (Bryc et al 2010).

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19, 1655-1664.

Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *PNAS* 107, 786-790.

Cheng CY, Reich D, Coresh J, Boerwinkle E, Patterson N, Li M et al. (2010) Admixture Mapping of Obesity-related Traits in african americans: The atherosclerosis Risk in communities (aRIc) Study. *Obesity* 18, 563–572.

Chow EJ, Puumala SE, Mueller BA, Carozza SE, Fox EE, Horel S et al. (2010) Childhood cancer in relation to parental race and ethnicity: a 5-state pooled analysis. *Cancer* 15, 3045-3053.

Hooker S, Hernandez W, Chen H, Robbins C, Benn Torres J, Ahaghotu C et al. (2009) Replication of Prostate Cancer Risk Loci on 8q24, 11q13, 17q12,19q33, and Xp11in African Americans. *The Prostate* 70, 270-275.

McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet* 5, 10.

Reed FA, Tishkoff SA (2006) African human diversity, origins and migrations. *Current Opinion in Genetics and Development* 16:597-605.

Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. *Am J Hum Genet* 8, 290–303.

Shlush LI, Bercovici S, Wasser WG, Yudkovsky G, Templeton A, Geiger D, Skorecki K (2010) Admixture mapping of end stage kidney disease genetic susceptibility using estimated mutual information ancestry informative markers. *BMC Med Genomics* 3, 47.

Shriner D, Adeyemo A, Chen G, Rotimi CN (2010) Practical considerations for imputation of untyped markers in admixed populations. *Genet Epidemiol* 34, 258-65.

Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual Admixture: analytical and study design considerations. *Genet Epidemiol* 28, 289-301.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 22, 1035-1044.

Winkler CA, Nelson GW, Smith MW (2010) Admixture mapping comes of age. *Genomics and Human Genetics* 11, 65–89.

Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D et al. (2011) Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nature Genetics* 43, 237–24.

Zheng W, Cai Q, Signorello LB, Long J, Hargreaves MK, Deming SL, Li G, Li C, Cui Y, Blo WJ (2009) Evaluation of 11 Breast Cancer Susceptibility Loci in African-American Women. *Cancer Epidemiol Biomarkers Prev* 18, 2761-2764.

CHAPTER IV

A specific Study: Replication and fine mapping of neuroblastoma SNP association at the *BARD1* locus in African-Americans

Latorre et al., Cancer Epidemiology, Biomarkers and Prevention (In Press)

Neuroblastoma (NBL) is a severe pediatric cancer with an incidence in white American children of approximately 11.5 per million (Stiller 1992). NBL is reported to be rarer among African-Americans, with an incidence of 8.5 per million (Stiller 1992). Limited data exist on incidence of NBL in Africa, but reports suggest it is similar to that of African-Americans or lower (Stiller 1992). Odds-ratio of NBL by parental ethnic origin relative to white American children was only 0.74 (95% CI 0.56-0.96) for children with both African-American parents (Chow 2010). However, a recent study has shown that black children in the US are more likely to have the high-risk form of the disease than white children (57% vs 44%; $P<0.001$), and have associated lower overall survival (OS) (67% vs 75%) and event free survival (EFS) (56% vs 67%) (Henderson 2010).

Genetic risk factors may contribute to disparities in cancer prevalence and outcome. In recent years, findings from case-control genome wide association studies (GWAS) (Maris 2008 , Capasso 2009, Diskin 2009, Wang 2010 , Nguyen 2011) and family-based linkage analysis (Mosse 2008) have improved our understanding of the genetic susceptibility to NBL.

GWAS have identified common variants within FLJ22536 (Maris 2008), BARD1 (Capasso 2009) and LMO1 (Wang 2010) as significantly associated with high-risk neuroblastoma. GWAS on low-risk cases identified SNPs within DUSP12, DDX4 and IL31RA (both at 5q11.2), and HSD17B12 as being associated with the less aggressive form of the disease (Nguyen 2011). Heritable mutations of ALK are the main cause of familial neuroblastoma (Mosse 2008).

Individuals of European descent constitute the majority of NBL patients in the United States, and so far genetic studies of NBL have been limited to this ethnic group. We have obtained genome-wide SNP data on a number of

African-American NBL patients collected by the Children's Oncology Group, and have used this information in a case-control study to evaluate whether the same genes and SNPs identified in the European-American studies affect risk of NBL in African-American children. Our results show that of all the risk variants identified so far, only SNPs of BARD1 unequivocally have an effect on NBL risk in African-American children. Whether this is due to difference in genetic susceptibility or limited power to detect small genetic effects remains to be determined.

Materials and Methods

Patients and controls

We performed an unmatched case control study to find association with NBL.

DNA samples and clinical information were available for 390 African-American NBL patients from the Children's Oncology Group. All subjects were African-Americans based on self-reported ethnicity. A total of 2500 control samples were selected based on self-reported African-American ethnicity from a large group of children collected by the Center for Applied Genomics at the Children's Hospital of Philadelphia (CHOP).

Genotyping

Genome-wide SNP genotype data from 390 NBL patients and 2500 disease-free control subjects were obtained using the Illumina HumanHap 550K (243 cases, 1875 controls) and Human610-Quad (147 cases, 625 controls) chips.

Only SNPs that were included in both chips were evaluated.

The Illumina Genotyping BeadChip is a relatively new method of performing multiplex SNP analysis. It consists of high density genotyping platforms
which enable whole genome genotyping of up to 655 thousand SNP markers. The large number of markers means that automated genotype calling procedures are required for assigning genotypes based on fluorescence intensities from hybridization.

Quality Control

Prior to statistical analysis, individual data were filtered on the basis of standard quality control measures, including call rates, discrepancy between reported sex and X chromosome marker heterozygosity, and cryptic relatedness, using the software PLINK. The software is able to rapidly manipulate and analyze large data sets comprising hundreds of thousands of markers genotyped for thousands of individuals with five main functions: data management, summary statistics, population stratification, association analysis, and identity-by-descent estimation.

A total of 38 samples (25 case subjects and 9 control subjects) had genotype yields of less than 95% and were removed, leaving 365 cases and 2491 controls available for analysis.

SNPs were excluded from further analysis if they showed deviation from Hardy–Weinberg equilibrium with a p-value less than 10^{-4} in controls or 10^{-7} in cases, a genotype yield of less than 95%, a minor allele frequency of less than 1%, and difference in missing rates between cases and controls with a p-value less than 10^{-4}. This filtering resulted in a total of 504,535 autosomal SNPs available for the subsequent analyses. Among these, we selected SNPs included in the genes showing association to NBL in previous GWAS of

European-American studies +/- an interval of 10 KB around them (Maris 2008 , Capasso 2009, Diskin 2009, Wang 2010 , Nguyen 2011), namely FLJ22536 (136 SNPs), BARD1 (25), LMO1 (30), DUSP12 (7), DDX4 (13), IL31RA (15), and HSD17B12 (27), for a total of 253 SNPs to be tested for association. To perform our analysis on African-Americans we took advantage of data from the International HapMap Project. Human genetic code discovery has made possible the development of a haplotype map of the human genome. The HapMap is a map of haplotype blocks and the specific SNPs that identify the haplotypes are called tag SNPs. In 10 million SNPs exist roughly 500,000 tag SNPs. Finding regions with genes that affect diseases is made much more efficient and comprehensive, since effort is not wasted typing more SNPs than necessary and all regions of the genome can be included.Coverage of common variants in the candidate genes was estimated using HapMap Phase 2 data in the CEU and YRI populations using the Tagger option of Haploview with pairwise tagging at r^2=0.8. Haploview is a software package that provides computation of linkage disequilibrium statistics and population haplotype patterns from primary genotype data. It generates marker quality statistics, LD information, haplotype blocks, population haplotype frequencies and single marker association statistics. The table shows how many of the SNPs in the dataset have been successfully tagged by the set of chosen tests. The mean r^2 represents the mean for only those SNPs successfully captured. Power to detect association in our sample for varying allele frequency and genetic risk effect was estimated using Quanto.

Analysis of stratification and admixture

African-Americans are a recently admixed population with two major founder groups. Admixture may introduce a bias in association tests if ancestry proportions are different between cases and controls. The availability of hundreds of thousands of markers across the genome allows for a very accurate empirical assessment of stratification due to admixture differences. Ancestry proportions in our cases and controls were estimated using the software ADMIXTURE and assuming 2 founder populations. ADMIXTURE is a tool for maximum likelihood estimation of individual ancestries that uses a block relaxation approach to alternately update allele frequency and ancestry fraction parameters.

We also evaluated stratification in our case-control sample by multidimensional scaling (MDS) using the procedure implemented in PLINK and a subset of independent SNPs (r^2<0.0001). PLINK provides a simple way to handle large GWAS data sets and assess confounding due to stratification and nonrandom genotyping failure. It performs classical multidimensional scaling (MDS) to visualize substructure, producing a k-dimensional representation, and provide quantitative indices of population genetic. One requirement of this approach is that SNPs are in approximate linkage equilibrium in the population. We pruned the SNP panel to a reduced subset of approximately independent SNPs using a repeated, sliding window procedure, recursively pruning SNPs based on pairwise r2. MDS data from our cases and controls were compared to those obtained from the HapMap Phase 2 in the 4 major populations (CEU, CHB, JPT, YRI).

Association analysis

To account for possible difference in substructure between cases and controls due to varying levels of admixture, we tested SNPs for association with NBL by logistic regression using the proportion of African ancestry estimated by ADMIXTURE as covariate. We also tested for association using a stratified Cochran-Mantel- Haenszel (CMH) test implemented in PLINK based on the clusters identified by the MDS analysis.

Then we performed association analysis for both chips separately, 550k and 610quad, and combined their p-values through a meta-analysis.

Results of the association tests with the three methods were almost identical and only results of logistic regression are reported.

To correct for multiple testing, we accounted for the number of SNPs tested in each gene by means of a Bonferroni correction, setting a threshold for significance equal to 0.05 divided by the number of SNPs tested in that particular gene. These correspond to 0.0004 for FLJ22536 (136 SNPs), 0.002 for BARD1 (25), 0.0017 for LMO1 (30), 0.0071 for DUSP12 (7), 0.0038 for DDX4 (13), 0.0033 for IL31RA (15), and 0.0019 for HSD17B12 (27). Since all these genes have been previously reported to be associated to NBL in multiple independent datasets, we did not correct for the number of genes tested. All p-values reported in our tables are uncorrected, asymptotic p-values from the corresponding tests.

Results

Power and SNP tagging

Our power calculation shows that we have greater than 80% power to replicate association at a nominal level of significance for MAF 10% or

more and Genetic Relative Risk 1.4 or more, which should be sufficient to replicate the association found in the Caucasian cohort (Figure 1a).
Percentage of HapMap Phase II SNP coverage for the individual genes provided by the SNPs included in our analysis varied from 50% for FLJ22536 to 75% for IL31RA in the YRI population, and from 70% for LMO1 to 91% for HSD17B12 in the CEU population (Figure 1b).

Stratification and admixture

MDS analysis of cases (Figure 2a) and controls (Figure 2b) together with the 4 major HapMap populations showed that our samples cluster along a continuum between the CEU and the YRI populations as expected .
Genome-wide estimates of African ancestry assuming 2 founder populations had a mean of 0.76%±0.23 in cases and 0.76%±0.19 in controls. There were some slight differences between the two groups with more cases than controls at the extremes of the distribution (0-10% and 80%-90% African ancestry) (Figure 2c).

Association analysis

Based on the genome-wide data, we estimated a genomic inflation factor (GIF) of the logistic regression test of 0.98. In contrast, the GIF of the 1 degree-of-freedom allelic test was 1.15. We are thus confident that the procedures implemented to control population stratification were effective in reducing any potential inflation in type 1 error. The GIF was obtained taking the median of the distribution of the chi-square statistic from logistic regression test results and dividing this median by the median of the corresponding (ideal) chi-square distribution. We expect this value to be close to one if there is no excess type 1 error due to possible bias.

Results for the most significant SNPs reported in previous GWAS of European-American patients are reported in Table 1. Two SNPs of BARD1, rs7587476 and rs6435862, showed significant p-values ($p<0.002$ when accounting for the number of SNPs tested in BARD1), and the other 3 BARD1 SNPs reached nominal significance. None of the SNPs in the other genes reached even nominal levels of significance (all p-values > 0.05). For the 5 BARD1 SNPs, the direction of the association was the same as the one observed in the European-American patients.

Because our previous studies in European-Americans had found that association signals were stronger when analysis was restricted to specific phenotypic categories (high-risk patients for FLJ22536, BARD1, and LMO1, and low-risk patients for DUSP12, DDX4, IL31RA, and HSD17B12), we repeated the association analysis in these two subgroups of patients (180 high-risk, 97 low-risk) separately against all controls. One SNP of BARD1 became more significant (rs7587476, $p=8x10^{-8}$), two had similar results to those observed in all patients, and two became not significant. One SNP in LMO1 reached nominal significance in the high-risk group (rs294938, p=0.03). All the other SNPs reported in previous studies were still not significant.

We then asked whether other candidate gene SNPs, different from those reported as most significant in the European-American patients, reached statistical significance in the African-American patients. Only one BARD1 SNP (in addition to those already reported in Table 1) showed significant association (rs16852804, p=0.0007). When we looked at the high-risk and low-risk subgroups separately, in the high-risk patients two additional BARD1 SNPs (rs7599060, p=0.0009, and rs16852804, p= 0.0018) and one SNP in LMO1 (rs4237769, $p=5x10^{-5}$) reached statistical significance. No

SNP showed statistical significance after correction for multiple SNPs tested in each gene in the low-risk group (all $p>0.01$).

Overall, three SNPs of BARD1 reached statistical significance after correction for the number of SNPs tested ($p<0.002$), and six more had p-values less than 0.05. To test whether these may represent multiple independent signals of association, we performed the logistic regression test conditional on the most significant SNP, rs7587476. The p-value for rs6435862 went from 0.00002 to 0.04, and for rs16852804 from 0.0007 to 0.005 (Figure 3). All other p-values were now greater than 0.05.

To examine the extent of the associated region, we plotted the r2 values relative to rs7587476 for all the SNPs in a 100Kb region around it against their genomic location. We chose rs7587476 as the reference SNP because not only is the most strongly NBL associated SNP in this study, but is also the most significant *BARD1* SNP in our most recent analysis of European Americans (8). Using data from the 1000 Genomes Project Pilot 1 in the CEU and YRI populations and the web- based software SNAP (15), we determined the size of the region which includes SNPs with r2 >0.5 with rs7587476 (Figure 4). In the YRI population, this region extends from 215,348,641 to 215,367,140 bp and comprises introns 2-4 and exons 3-4 of BARD1. In the CEU population, this region extends from 215,344,039 to 215,457,501 bp, and thus for an additional 4.6 Kb proximally and 90.4 Kb distally.

Discussion

SNPs of BARD1 associated to NBL in European-American patients show similar, strongly significant association in African-Americans. In particular

most of the association detected in BARD1 seems to be explained by one SNP, rs7587476, located in intron 3, with some residual association signal detected by two other SNPs located in the first intron. In contrast, the association in European Americans extended further to intron 4, possibly due to the more extensive LD around the associated SNPs. This information may be helpful in locating the causal BARD1 variants.

Besides the BARD1 SNPs, the only other SNP significant after correction for number of SNPs tested in a given gene was rs4237769 ($p=5x10^{-5}$) in LMO1, when analysis was restricted to patients from the high-risk group. Five other SNPs in or around LMO1 had p-values <0.05 in the high-risk group, including rs204938 (p=0.04) reported associated to NBL by Wang et al (2011) The second most significant LMO1 SNP in the high-risk group in this study, rs3794012 (p=0.005) also had a p-value of $3x10^{-5}$ in Wang et al (2011). However, the rare C allele is associated to NBL in the African-American patients, rather than the common T allele as observed in the European American cases.

The different structure and lower linkage disequilibrium in the African-American population may have prevented us to detect association in the other genes.

Interestingly we did not detect association with SNPs of FLJ22536, which are the most strongly associated to NBL in European-Americans (Wang). The most significant FLJ22536 SNP in our study was rs1207774 with a p-value of 0.005 in the high-risk group. However this is hardly significant when considering the large number of SNPs tested in this gene (136). Based on HapMap CEU and YRI Phase II data, among the genes tested coverage of variation in FLJ22536 is the lowest in Africans (50%) and the most divergent from that of European-Americans (78%).

None of the genes associated to the low-risk group in European-Americans (ref) showed significant association in our study, either in all patients or in the low-risk sub-category, after accounting for the multiple SNPs tested. However this risk group includes only 97 African-American cases and limited power may have prevented us to detect a significant association.

This study shows the difficulty to detect association in African-Americans even for SNPs that are confirmed and show strong significant association in multiple European or European-American populations. A more detailed analysis genotyping additional SNPs not included in the GWAS chip will be necessary to be able to conclude on the role of these genes on susceptibility to NBL in African-Americans. Furthermore, a larger sample of African American patients would allow genome-wide analysis and possibly detection of novel association.

Table 1. Results of logistic regression test of association in African-American NBL patients for most significant SNPs in European-American GWAS

Gene	Chr.	SNP (bp)	Alleles (m/M)	MAF cases	MAF controls	p-value	OR	95% CI
FLJ22536	6	rs4712653 (22233943)	T/C	0.19	0.19	0.56	0.94	0.77-1.15
		rs9295536 (22239908)	C/A	0.19	0.19	0.30	0.90	0.73-1.10
		rs6939340 (22247983)	A/G	0.18	0.17	0.77	0.97	0.79-1.19
*BARD1**	2	rs3768716 (215344039)	G/A	0.10	0.07	0.007	1.45	1.11-1.90
		rs17487792 (215351745)	T/C	0.10	0.06	0.009	1.45	1.10-1.93
		rs7587476 (215362132)	T/C	0.42	0.34	$2.4x10^{-6}$	1.47	1.25-1.72
		rs6712055 (215375149)	C/T	0.22	0.17	0.006	1.31	1.08-1.59
		rs6435862 (215380791)	G/T	0.34	0.26	$1.8x10^{-5}$	1.44	1.22-1.70
LMO1	11	rs4758051 (8195215)	A/G	0.20	0.18	0.77	1.03	0.84-1.27
		rs10840002 (8199602)	G/A	0.27	0.25	0.79	1.03	0.85-1.23
		rs110419 (8209429)	G/A	0.22	0.23	0.18	0.88	0.72-1.07
		rs204938 (8234773)	T/C	0.34	0.31	0.27	1.10	0.93-1.30

*SNP rs6715570 reported by Capasso et al. (6) was removed from our analysis for missing rate greater than 5% in cases.

m/M = minor allele/major allele

MAF = minor allele frequency

CI = confidence interval

Figure legends

Figure 1.

Power and tagging analysis. a) Power to detect association for varying odds-ratios (1.2-1.6) and risk allele frequencies (0.1- 0.9) at significance levels defined by a Bonferroni correction for the number of SNPs tested in the candidate genes (0.0004 for *FLJ22536*; 0.0016 for *LMO1*; 0.0015 for *BARD1* (not shown) was identical to power for *LMO1*). b) Percentage of HapMap Phase II SNPs in the three candidate genes in the YRI and CEU populations tagged by the SNPs included in the analysis.

Figure 2.

MDS plots and distribution of African ancestry. Multidimensional scaling plots of cases (a) and controls (b) against the four major HapMap populations. c) Distribution of percentage of African ancestry in cases and controls.

Figure 3.

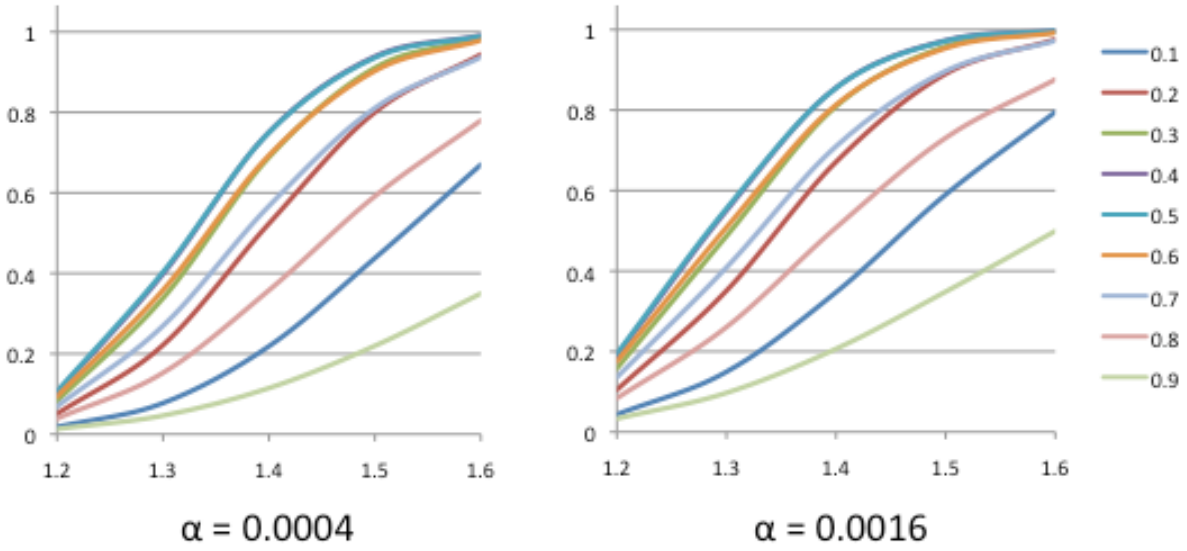
Regional association plots for *BARD1* SNPs. a) Negative log10 p-values from logistic regression analysis for all SNPs tested in the *BARD1* region plotted against their genomic location. b) Negative log10 p-values from logistic regression analysis conditional on association at rs7587476. Color shading for r2 is relative to the most significant SNP (rs7587476) based on data from the YRI population.

Figure 4.

Regional LD plots in *BARD1* genomic region. r2 values relative to rs7587476 in the YRI (a) and CEU (b) populations for SNPs in a 100Kb window around it plotted against their genomic location. Data are from the 1000 Genomes Project Pilot 1. Dotted vertical lines delimit regions including SNPs with r2>0.5.

Figure 1

a



α = 0.0004 α = 0.0016

b

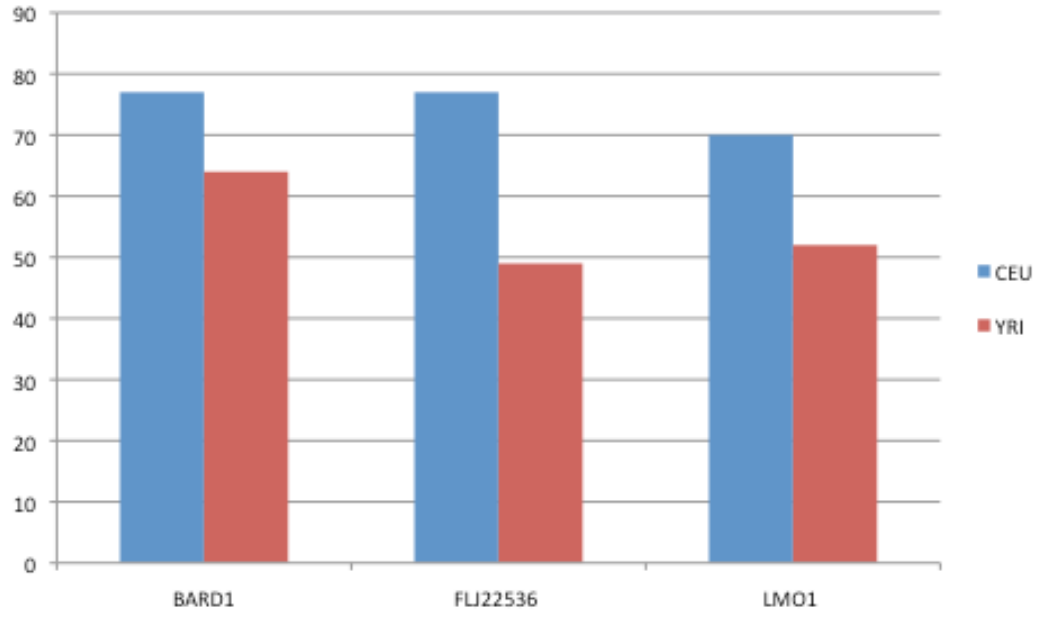
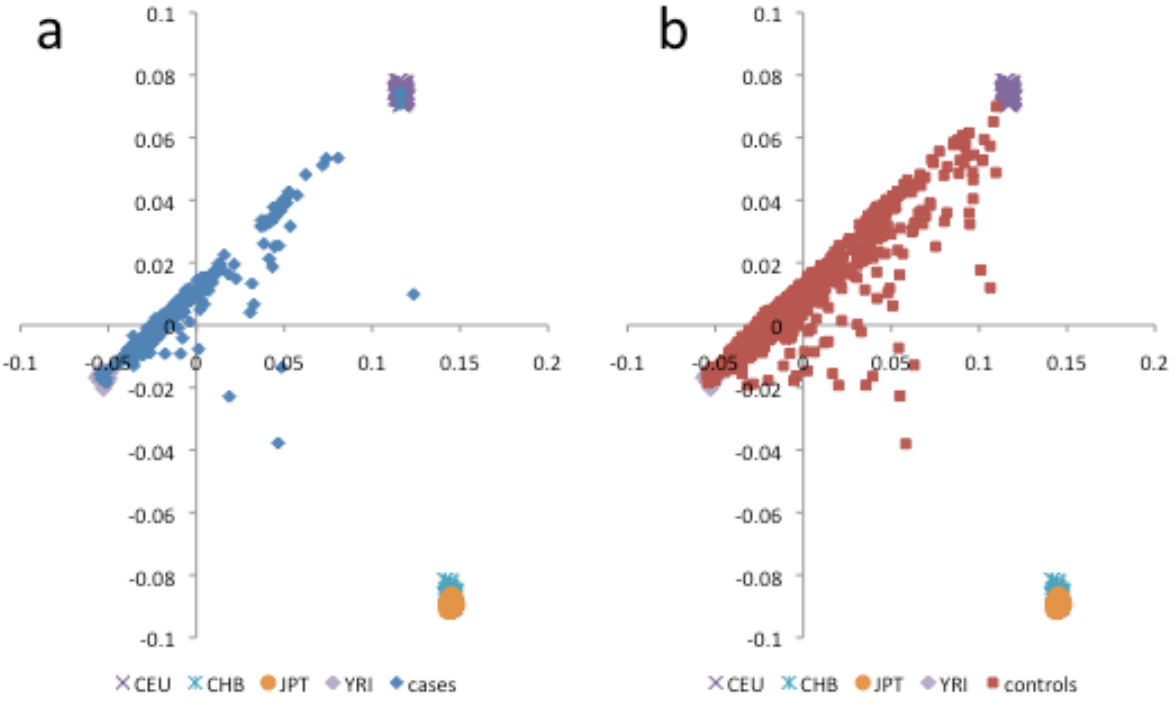


Figure2


a
b
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.1
-0.05
0
0.05
0.1
0.15
0.2
CEU
CHB
JPT
YRI
cases
controls

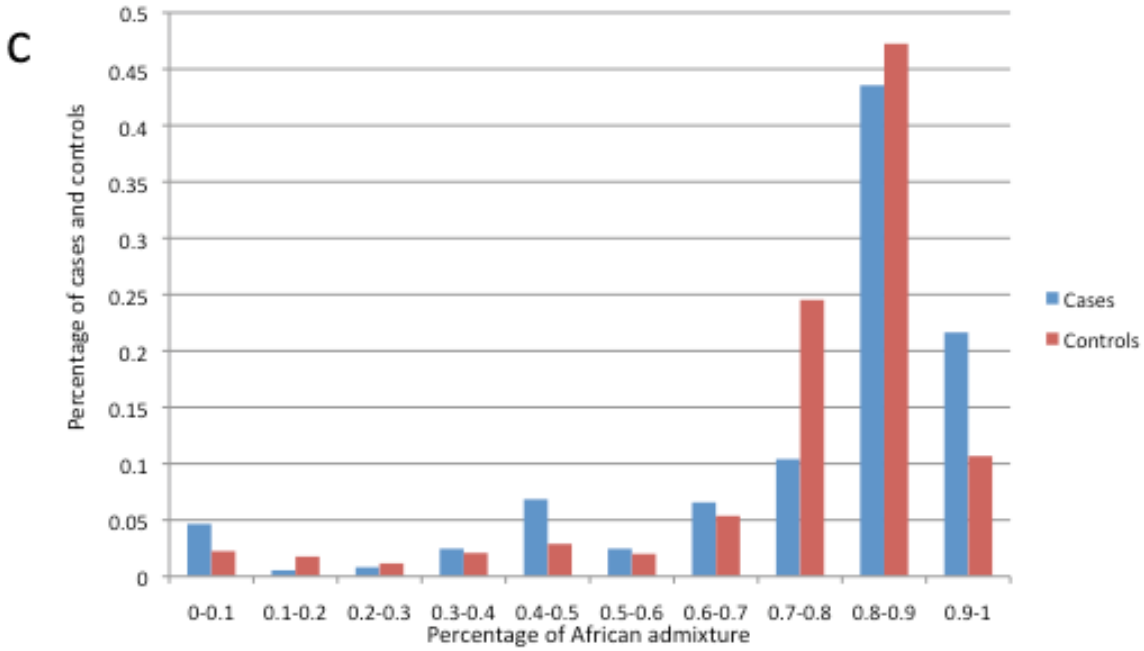
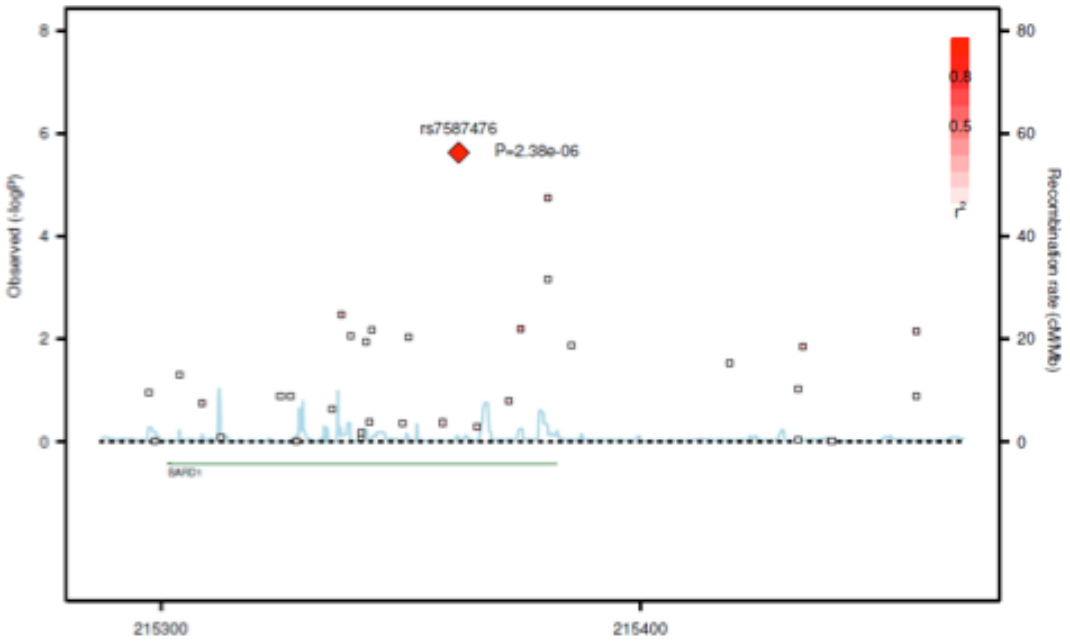

c
Percentage of cases and controls
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
Percentage of African admixture
Cases
Controls

Figure 3

a


rs7587476
P=2.38e-06
Observed (-logP)
Recombination rate (cM/Mb)
215300
215400

b

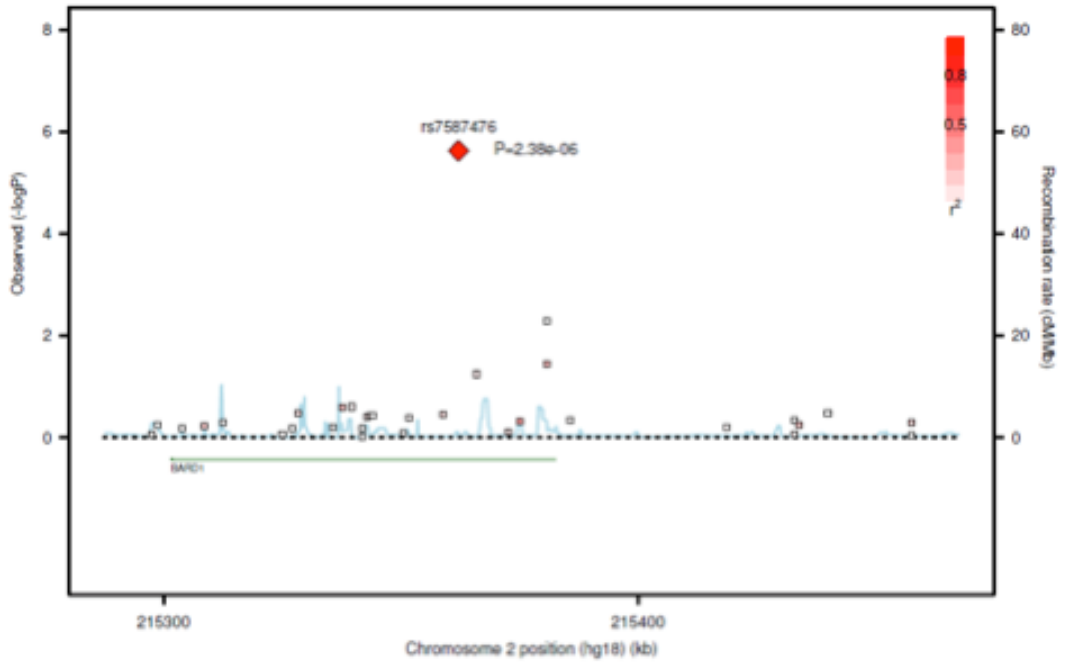
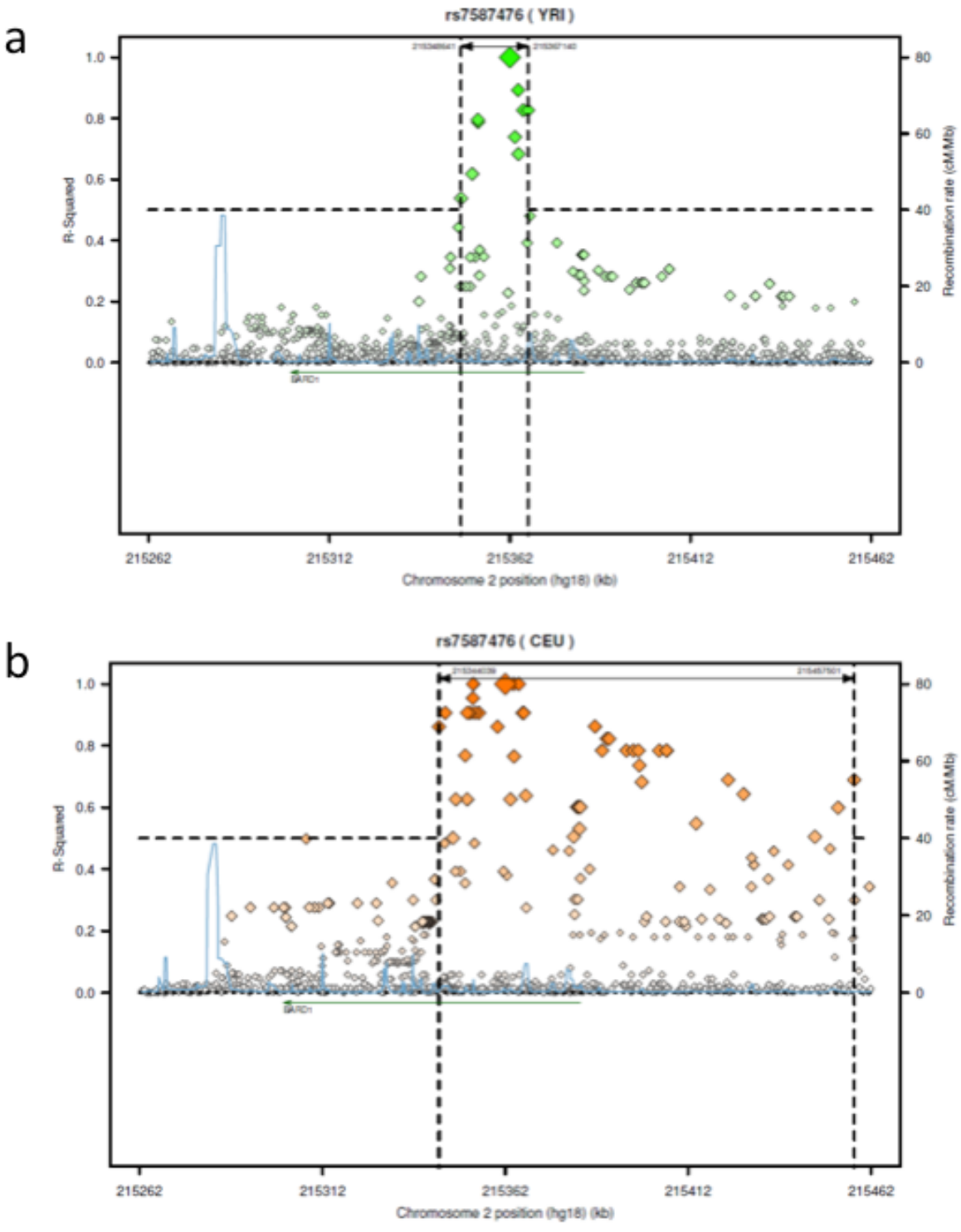

rs7587476
P=2.38e-06
Observed (-logP)
Recombination rate (cM/Mb)
215300
215400
Chromosome 2 position (hg18) (kb)

Figure 4


a
rs7587476 (YRI)
R-Squared
Recombination rate (cM/Mb)
Chromosome 2 position (hg18) (kb)
215262
215312
215362
215412
215462
b
rs7587476 (CEU)
R-Squared
Recombination rate (cM/Mb)
Chromosome 2 position (hg18) (kb)
215262
215312
215362
215412
215462

Methodological Appendix

Hardy Weinberg Equilibrium

Genotype frequencies at any locus are a simple function of allele frequencies in absence of migration, mutation, natural selection, and assortative mating. This phenomenon is called "Hardy-Weinberg equilibrium". We used exact tests for HWE for our large-scale study of SNP data including hundreds of thousands of markers.

Considering a sample of SNP genotypes for *N* unrelated diploid individuals measured at an autosomal locus, the sample includes 2*N* alleles, with *nA* copies of the A rarer allele and *nB* copies of the common allele. Let the number of heterozygous AB genotypes be *nAB* and the numbers of AA and BB homozygous genotypes are $n_{AA} = (n_A - n_{AB})/2$ and $n_{BB} = (n_B - n_{AB})/2$.

There are $(2N)!/n_A!n_B!$ possible arrangements for the alleles in the sample and $2^{n_{AB}} N!/(n_{AA}!n_{AB}!n_{BB}!)$ of these arrangements correspond to exactly n_{AB} heterozygotes.

Thus, the probability of observing exactly n_{AB} heterozygotes in a sample of *N* individuals with n_A minor alleles under the assumption of HWE is

$$P(N_{AB} = n_{AB} \mid N, n_A) = \frac{2^{n_{AB}} N!}{n_{AA}!n_{AB}!n_{BB}!} \times \frac{n_A!n_B!}{(2N)!} . \quad (1)$$

Equation (1) holds for each possible number of heterozygotes, n_{AB}. When n_A is odd, possible numbers of heterozygotes are 1, 3, 5,…, n_A and when n_A is even they are 0, 2, 4,…, n_A. The expression for $P(n_{AB} \mid N, n_A)$ given in equation (1) leads to natural tests for HWE.

Hapmap Project

The HapMap is an international project that aims to identify and catalog genetic similarities and differences in human beings comparing the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared. The ultimate goal is to help find genes that affect health, disease, and individual responses to medications and environmental factors. Over the course of many generations, segments of the ancestral chromosomes in an interbreeding population are shuffled through repeated recombination. They are separated by places where recombination has occurred but they have not been broken up by it. These segments are the haplotypes, a group of genes within an organism that was inherited together from a single parent because of genetic linkage, the tendency of certain loci or alleles to be inherited together. In addition, the term "haplotype" can also refer to the inheritance of a cluster of single nucleotide polymorphisms (SNPs), which are variations at single positions in the DNA sequence occurring when a single nucleotide — A, T, C, G --- in the genome differs between members of a biological species or paired chromosomes in an individual. By examining haplotypes it is possible to identify patterns of genetic variation that are associated with health and disease states analyzing stretches of DNA near the SNP cluster to try to

identify the genes responsible for causing the disease. The chromosomes in human cells occur in pairs (with the exception of the sex cells) of exactly 46 chromosomes, 22 homologous pairs of autosomes and one pair of sex chromosomes. The complete set is the diploid complement where one member of each chromosome pair is inherited from the father and the other from the mother. The two homologues have the same sequence of genes in the same positions, but they can be distinguished because of sequence variations at several loci. The haplotypes in the human genome have been produced by the molecular mechanisms of sexual reproduction and by the history of our species. It might seem that there is a 50% probability that any given gamete receives one chromosome rather than the other from a particular homologous pair, and that there are 2 to the power of 23 distinct gametes that any given individual might produce. Instead, when sperm and egg cells are being formed the gamete receives a mixture of the two homologous chromosomes because of crossover. Crossover can split alleles that lie together on a common parental chromosome and results in a hybrid chromosome containing pieces from both members of a chromosome pair that could contains alleles that originally came from different grandparents. This process is called recombination; the further apart two genes are, the higher the probability of an odd number of crossovers and therefore of observing a recombination, to a maximum of 50% frequency. The proportion of meioses that result in a recombination is called “recombination fraction” and it is an indication of how far apart two genes are. From a genetic point of view it seems that all humans today are descended from anatomically modern ancestors who lived in Africa about 150,000 years ago. Humans migrated out of Africa carrying with them part but not all of the genetic variation of the ancestral population. Thus, the haplotypes in non-

African populations tend to be subsets of the haplotypes inside Africa and tend to be longer because populations in Africa have been larger through much of our history and recombination has had more time there to break up haplotypes. Through random chance, natural selection, and other genetic mechanisms, the frequency of haplotypes came to vary from region to region as modern humans spread throughout the world. A given haplotype can occur at different frequencies in different populations, especially when those populations are widely separated and unlikely to exchange much DNA through mating. A gene mutation is a permanent change in the DNA sequence that makes up a gene. Gene mutations occur in two ways: they can be inherited from a parent or acquired during a person's lifetime, when it is caused by environmental factors such as ultraviolet radiation from the sun, or if a mistake is made as DNA copies itself during cell division. Mutations create new haplotypes, and most of the recently arising haplotypes have not had enough time to spread widely beyond the population and geographic region in which they originated. Because of the history of the human species, most of the common haplotypes in human chromosomes occur in all human populations, some haplotypes may be more common in one population and less common in another, and newer haplotypes may be found in just a single population.

The 4 HapMap populations included in our analysis are:

- Yoruba in Ibadan, Nigeria (YRI): These samples were collected in a particular community in Ibadan, Nigeria, from individuals who identified themselves as having four Yoruba grandparents. These samples should not be described merely as "African," "Sub-Saharan

African," "West African," or "Nigerian," since each of those designators encompasses many populations with many different ancestral geographies.

- Japanese in Tokyo, Japan (JPT): These samples were collected in the Tokyo metropolitan area, from people who came from (or whose ancestors came from) many different parts of Japan. Thus, this set of samples can be viewed as representative of the majority population in Japan.
- Han Chinese in Beijing, China (CHB): These samples were collected from individuals living in the residential community at Beijing Normal University who were self-identified as having at least three out of four Han Chinese grandparents. Although individuals of Beijing Normal University were from many different parts of China, this set of samples was not drawn to be representative of all Han Chinese people. The samples also should not be seen as representing all people in China, where there are 56 officially recognized ethnicities.
- CEPH (Utah residents with ancestry from northern and western Europe) (CEU): These samples were collected from people living in Utah with ancestry from northern and western Europe. The term "CEPH" stands for the Centre d'Etude du Polymorphisme Humain, the organization that collected these samples in 1980. Because the importance of precision in assigning group membership to prospective donors based on ancestral geography was not well appreciated in 1980, it is unclear how accurately these samples reflect the patterns of genetic variation in people with northern and western European ancestry.

Power analysis

QUANTO allows for unequal numbers of cases and controls in the sample, with K denoting the control-to-case ratio.

The likelihood for the logistic model for N cases and N×K controls has form

$$L(\alpha,\beta_g,\beta_e,\beta_{ge}) = \prod_{i=1}^{N} \frac{e^{\alpha+\beta_g G_i+\beta_e E_i+\beta_{ge} G_i E_i}}{1+e^{\alpha+\beta_g G_i+\beta_e E_i+\beta_{ge} G_i E_i}} \prod_{j=1}^{N\times K} \frac{1}{1+e^{\alpha+\beta_g G_i+\beta_e E_i+\beta_{ge} G_j E_j}}$$

where the first product is taken over the N cases and the second is taken over the N×K controls. MLE's obtained from this model are consistent estimators of the log-odds-ratio parameters from the logistic model.

Admixture

Population stratification is a confounding factor in genetic association studies. To correct its effects we used an approach known as "structured association" and estimated ancestries from the genotypes actually collected in our study through the software ADMIXTURE.

We estimated the global ancestry that is the proportion of ancestry from each contributing population, considered as an average over the individual's entire genome, while in the local ancestry paradigm each person's genome is divided into chromosome segments of definite ancestral origin. "Global ancestry estimation" consider two approaches: "model-based ancestry estimation" and "algorithmic ancestry estimation".

ADMIXTURE estimates ancestry coefficients as the parameters of a

statistical model, and simultaneously population allele frequencies along with ancestry proportions. It performs a block relaxation algorithm that alternates between updating the ancestry coefficient matrix Q and the population allele frequency matrix F. Finally, there is an acceleration of the convergence of block relaxation by a novel quasi- Newton method. Here we present the underlying statistical model and describe the optimization techniques used to maximize the likelihood.

Genotype data consist of a large number J of SNPs from a large number I of unrelated individuals from an admixed population with contributions from K postulated ancestral populations.

Population k contributes a fraction qik of individual i's genome. Allele 1 at SNP j has frequency fkj in population k, and both the qik and the fkj are unknown. We are primarily interested in estimating the qik to control for ancestry in an association study, but our approach also yields estimates of the fkj.

The model makes the assumption of linkage equilibrium among the markers and our SNP set was pruned to mitigate background linkage disequilibrium (LD).

Let gij represent the observed number of copies of allele 1 at marker j of person i. Thus, gij equals 2, 1, or 0 accordingly, as i has genotype 1/1, 1/2, or 2/2 at marker j. Since individuals are considered independent, the log-likelihood of the entire sample is

$$L(Q,F)=\sum_i\sum_j\left\{g_{ij}\ln\left[\sum_k q_{ik}\right]+(2-g_{ij})\ln\left[\sum_k q_{ik}(1-f_{kj})\right]\right\}. \quad (2)$$

up to an additive constant that does not enter into the maximization problem.

The parameter matrices Q = {qik} and F = {fkj} have dimensions I x K and K x J, for a total of K(I + J) parameters. As an optimization method the EM algorithm is used to get quickly to the vicinity of the maximum and then shift to accelerated block relaxation that alternates updates of the Q and q parameters.

References

Capasso M, Devoto M, Hou C, Asgharzadeh S, Glessner JT, Attiyeh EF, et al. (2009). Common variations in *BARD1* influence susceptibility to high-risk neuroblastoma. *Nat Genet* 41, 718-723.

Chow EJ, Puumala SE, Mueller BA, Carozza SE, Fox EE, Horel S, et al. (2010). Childhood cancer in relation to parental race and ethnicity: a 5-state pooled analysis. *Cancer* 116, 3045- 3053.

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, et al. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459, 987-991.

Henderson TO, Bhatia S, Pinto N, London WB, McGrady P, Crotty C, et al. (2011). Racial and ethnic disparities in risk and survival in children with neuroblastoma: a Children's Oncology Group study. *J Clin Oncol* 29, 76-82.

Maris JM (2010). Recent advances in neuroblastoma. *N Engl J Med* 362, 2202–2211

Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH, et al. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N Engl J Med* 358, 2585-2593.

Nguyen LB, Diskin SJ, Capasso M, Wang K, Diamond MA, Glessner J, et al. (2011) Phenotype restricted genome-wide association study using a gene-

centric approach identifies three low-risk neuroblastoma susceptibility loci. PLoS Genet 7, e1002026.

Stiller CA and Parkin DM (1992). International variations in the incidence of neuroblastoma. *Int J Cancer* 52, 538-543.

Wang K, Diskin SJ, Zhang H, Attiyeh EF, Winter C, Hou C, et al. (2011) Integrative genomics identifies *LMO1* as a neuroblastoma oncogene. *Nature* 469, 216-220.

Aknowledgements

I would like to thank my Supervisor, Professor Giuseppe Passarino, and all the laboratory of Genetics at the University of Calabria.
A special thank goes to Professor Marcella Devoto. I worked at her side for almost two years and her honesty, professionality and respect inspired me to work at my best.
I'm grateful to Matt for correcting my written and spoken English every time it was needed.
At last but not least, I can't not say thanks to my family (Gianni, Flora, Andrea e Anna), Silvio, Ada, Silvia, Miryung, Pino and Concettina for supporting me and for showing me their love.

APPENDIX

Montesanto A, Latorre V, Giordano M, Martino C, Domma F, Passarino G: **The genetic component of human longevity: analysis of the survival advantage of parents and siblings of Italian nonagenarians.** Eur J Hum Genet. 2011; 19:882-886.. Epub 2011 Mar 16.

European Journal of Human Genetics (2011), 1–5

www.nature.com/ejhg

ARTICLE

The genetic component of human longevity: analysis of the survival advantage of parents and siblings of Italian nonagenarians

Alberto Montesanto[1], Valeria Latorre[1], Marco Giordano[1], Cinzia Martino[1], Filippo Domma[2] and Giuseppe Passarino*[,1]

Many epidemiological studies have shown that parents, siblings and offspring of long-lived subjects have a significant survival advantage when compared with the general population. However, how much of this reported advantage is due to common genetic factors or to a shared environment remains to be resolved.We reconstructed 202 families of nonagenarians from a population of southern Italy. To estimate the familiarity of human longevity, we compared survival data of parents and siblings of long-lived subjects to that of appropriate Italian birth cohorts. Then, to estimate the genetic component of longevity while minimizing the variability due to environment factors, we compared the survival functions of nonagenarians' siblings with those of their spouses (intrafamily control group).We found that both parents and siblings of the probands had a significant survival advantage over their Italian birth cohort counterparts. On the other hand, although a substantial survival advantage was observed in male siblings of probands with respect to the male intrafamily control group, female siblings did not show a similar advantage. In addition, we observed that the presence of a male nonagenarians in a family significantly decreased the instant mortality rate throughout lifetime for all the siblings; in the case of a female nonagenarians such an advantage persisted only for her male siblings.The methodological approach used here allowed us to distinguish the effects of environmental and genetic factors on human longevity. Our results suggest that genetic factors in males have a higher impact than in females on attaining longevity.
European Journal of Human Genetics advance online publication, 16 March 2011; doi:10.1038/ejhg.2011.40

Keywords: human longevity; genetic component; inheritance; familial determinants

INTRODUCTION

In the last decade many epidemiological studies on human longevity have shown that parents, siblings and offspring of long-lived subjects have a significant survival advantage compared with the general population in attaining longevity.[1–11] Although these studies do not distinguish between shared environmental and genetic factors, twin data suggest that genes may have a modest role in achieving longevity.[12,13] In order to better distinguish the effect of genes from the effect of shared familial environment, Schoenmaker *et al*[3] analyzed the survival data of the spouses of long-lived subjects as an additional control group. They found that members of this control group, who shared most of their adult life with the long-lived partner, did not show any advantage/benefit in terms of survival, suggesting that a substantial contribution in the familiarity of human longevity is attributable to genetic factors. However, as a complex trait, the heritability of 'lifespan' may be influenced by an interplay of genetic, environmental and stochastic factors.[14,15] In addition, the influence of the genetic component on lifespan is expected to be stronger in populations of areas where environmental factors are harsher[16] as demonstrated in different studies.[9,17,18]

Calabria is one of the poorest Italian regions located in the southern part of the peninsula. In the present study we aimed (i) to estimate the familial component of human longevity in Calabrian population; (ii) to uncouple within such a familial component the genetic from the environmental component. For these purposes, we reconstructed 202 pedigrees of Calabrian families where at least one nonagenarian individual was present. In order to estimate the presence of a familial component of longevity, we compared the survival data of parents and siblings of long-lived subjects with appropriate Italian birth cohorts. Then, to minimize the variability of familial environmental factors, we compared the survival functions of long-lived siblings with those of their spouses (intrafamily control group). This approach allowed us to estimate how much of the familiarity of the analyzed phenotype is due to genetics.

MATERIALS AND METHODS

Our sample consisted of the members of 202 families identified in seven municipalities (Bisignano, Cariati, Cosenza, Luzzi, Montalto Uffugo, Rende, and Rose) of Calabria (southern Italy). Each municipality was contacted in 2006 and invited to send a list of subjects living in their territory born in 1916 or before (probands). In total, 1475 eligible probands were identified. In the present study, which started in October 2008, we reconstructed the family pedigree of 202 probands.

Age validation

For complete age validation of long-lived individuals, their parents, siblings and the long-lived spouses of siblings, the following documents were

[1]Department of Cell Biology, University of Calabria, Rende, Italy; [2]Department of Economics and Statistics, University of Calabria, Rende, Italy
*Correspondence: Professor G Passarino, Department of Cell Biology, University of Calabria, 87036, Rende, Italy. Tel: +39 0984 492932; Fax: +39 0984 492911; E-mail: g.passarino@unical.it
Received 4 November 2010; revised 8 February 2011; accepted 9 February 2011

examined: the birth certificate, marriage certificate(s), the population registry (*Anagrafe*) personal sheet, the birth certificate of both parents (except for non-related parents) and death certificates. In addition, in order to further confirm the completeness of the reconstructed pedigrees, specialized personnel contacted a relative (usually a child or nephew/niece) for each proband whose genealogical tree had been reconstructed to verify information regarding the name(s), places and dates of birth, marriage, death, and emigration of the parents, all siblings and their spouses of the long-living probands.

Statistical analyses

As we focused on the mortality and survivorship of their parents and siblings, the probands were not included in the analysis described here. Because of our interest in longevity, we examined the survival patterns of the parents and siblings of the probands conditional on survival to age 30. We chose age 30 as a cutoff because siblings who died at younger ages probably did so because of stochastic, non-heritable factors (eg, infectious diseases, accidents, violence).[4,9] This minimizes the effect of such errors on cumulative survival probability.

In order to verify whether parents of probands lived longer than expected, we compared their life span with that of their respective Italian birth cohort. We first estimated the mean age at death of proband's parents conditional on survival to age 30. All parents had died, and thus their survival experience was complete. Following Perls *et al*,[19] we then matched each participant by year of birth and sex with their respective Italian cohort to obtain life expectancies conditional on survival to age 30. For the Italian population, sex-specific life tables are available from the Human Mortality Database (HMD) with the percentages of death for each year of age in the range of 0–100 years and each birth year since 1872 (http://www.mortality.org). The weighted average of these cohort-specific estimates was then compared with the corresponding estimates obtained for the parents of the probands.

Death rates for siblings and their spouses were computed, separately, from tabulations by age of sibling deaths and censored observations. Both the death counts and exposure estimates were aggregated in 5-year age groups. Standard demographic methods were used to calculate the mortality rate and its variance. Death rates, d_x, were computed as the ratio of deaths, D_x, over the exposure-to-risk E_x in a given age group:

$$d_x = \frac{D_x}{E_x}$$

E_x was calculated as the number of sibling survivors at the beginning of an age interval, N_x, minus half of the deaths, D_x, and censorings, W_x, during the interval:

$$E_x = N_x - \frac{1}{2}(D_x + W_x)$$

The variance of the estimated mortality rate was calculated according to Poisson distribution.[20]

The survival rate for interval x was computed as following:

$$p_x = \frac{R_x - D_x}{R_x}$$

The risk-set R_x equaled the number of sibling survivors at the beginning of an age interval, minus half of the censorings over that interval:

$$R_x = N_x - \frac{1}{2}W_x$$

The survival curves, S_x, were computed as $S_x = p_0 p_1 \ldots p_{x-1}$.

Standard errors for sibling survival probabilities were calculated based on an assumption of binomial variability (conditional on the observed collection of R_x values) using Greenwood's formula.[21] The obtained survival curves were then compared by log-rank test.

In order to investigate whether proband siblings had lower mortality and higher probability of surviving at advanced ages, siblings survival curves were compared with (i) the corresponding survival curves of the 1910 birth cohort for the general Italian population (the average year of birth for siblings was 1911) and (ii) the survival curves of their spouses (intrafamily control group). In this case, as survival experiences of proband siblings were not complete (some were still alive at the time of the study) the approach used for the parents of the probands was not applicable. To bypass this problem, we used an approach widely applied in other studies[1,9] that is, to define a 'control group' by determining the mean year of birth of the siblings of the probands. Then, we compared their survival experience with respect to those of the Italian birth cohort of such year. Survival data from the 1910 cohort were derived from the HMD. As in the previous case, survival probabilities were conditional on survival to age 30. The siblings of the probands and their respective spouses who emigrated from Italy were excluded from the study and their immigration periods were used as censoring dates. The exclusion circumvented the introduction of a bias due to the effect exerted on the phenotype by the 'new' environment in which they went to live.

In order to quantify the survival advantage due to a presence of a long-lived individual in the reconstructed family, the siblings' hazard function was compared with those of their spouses using a Cox regression model.[22] In this model 'relationship to the proband', 'gender of the sibling/spouse' and their interaction were used as explanatory covariates.

RESULTS

Table 1 reports a descriptive analysis of the subjects analyzed for this study. Of the 202 probands (126 women and 76 men), 129 were deceased (63.9%) at the time of this analysis and 73 (36.1%) were alive. The probands had a median of six siblings with a range of 1–13. A total of 1160 siblings, 593 men and 567 women, were identified for the analysis. Of these, 90 (15.2%) males and 105 (18.5%) females died in childhood (0–10 years of age). Of the remaining, at the time of data collection 63 (12.5%) male and 68 (14.7%) female siblings were still alive. These and siblings who migrated produced a total of 179 (18.5%) censored observations. In addition, a total of 669 non-related individuals (spouses of siblings, 298 men and 371 women) were identified for the same analysis. At the time of data collection, 18 (6.0%) male spouses and 90 (24.3%) female spouses were still alive. These and siblings' spouses who migrated outside of Italy gave a total of 128 (19.1%) censored observations. In the case of the siblings, early childhood mortality was included, hence the relatively large difference in number of deceased *vs* deceased ≥ 30.

The average year of birth of the probands was 1910 and for their siblings the average was 1911. With regard to parent's data, for 43 mothers and 33 fathers information on age at death were unknown. The average year of birth for fathers was 1876 and for mothers 1882.

Median ages at death for fathers and mothers of the probands were 77.5 and 79 years, respectively. Excluding deaths which occurred before age 30, the median age at death of the siblings of probands was higher than those observed in the relevant spouses (78 years in male siblings of probands *vs* 75 of the male spouses; 81 years in female siblings of probands *vs* 79 of the female spouses).

Table 2 shows the results for comparisons of mean ages at death of the proband's parents with the corresponding estimates for Italian birth cohort conditioned on survival to the age of 30 years. The mean age at death of the father's of probands was about 75 years. These estimates were substantially higher than the corresponding estimates for the respective Italian birth cohorts. In fact, the mean age at death was about 11% higher (8.05 years, $P<0.001$) when compared with the relevant Italian birth cohorts.

Figure 1 shows the survival curves obtained for the siblings of the probands and the 1910 Italian birth cohort. Both curves are conditioned for survival to the age of 30 years, as reported in Materials and Methods. Although the 1910 Italian birth cohort is not totally extinguished, Figure 1 shows the presence of a substantial survival advantage, which is more evident in male siblings ($P<0.001$) than in females ($P=0.01$).

Table 1 Characteristics of the subjects (belonging to 202 families) analyzed in the study. Median age and interquartile range are displayed

	Men		*Women*	
	N	*Age*	*N*	*Age*
Parents of proband				
Deceased	167[a]	77.5 (68–85)	157[b]	79 (70–86)
Deceased (≥30 years)	166	78 (68–85)	157	79 (70–86)
Sibling				
Alive	63[c]	85 (81–89)	68[c]	86 (81–90)
Deceased	450	73 (33.75–82)	481	74 (22.5–85)
Deceased (≥30 years)	384	78 (68–84)	355	81 (71–86)
Proband sibling's spouses[d]				
Alive	18[c]	84 (82–87.5)	90[c]	82.5 (77–87)
Deceased	272	75 (64–82)	269	79 (71–85)
Deceased (≥30 years)	270	75 (64.75–82)	267	79 (71–85)
1910 Italian birth cohort				
Life expectancy at birth		49.33		54.52
Life expectancy conditional on survival to age 30 years		71.12		78.15

[a]33 fathers had unknown age at death.
[b]43 mothers had unknown age at death.
[c]Censored for immigration not included.
[d]Calculations include only the first spouse.

Table 2 Comparisons of mean ages at death (SE in parenthesis) conditioned on survival to age 30 of parents of probands with the respective Italian birth cohort, birth years 1876 for fathers, 1882 for mothers

Parents of probands	*Mean age at death by sex conditional on survival to age 30*	*Italian cohort life tables*[a]	*Excess years*	*P-value*[b]
Men (*N*=166)	75.49 (0.98)	67.44	8.05	<0.001
Women (*N*=157)	76.06 (1.10)	70.78	5.28	<0.001

Note: source, Human Mortality Database: http://www.mortality.org; calculations by the authors.
[a]Life expectancy at birth conditioned on survival to age 30 for the Italian birth cohort. Cohort life table estimates were assumed to have zero variance.
[b]*P*-value refers to *t*-Student's test.

Figure 2 compares the survival curves of the siblings of probands with those of their spouses. A substantial survival advantage is observed in male siblings of probands with respect to the male spouses ($P<0.001$). This is not true for women ($P=0.950$). In both genders the chances of survival for the two groups does not differ substantially during early adulthood. However, after the age of 50, the survival patterns begin to diverge in favor of male siblings with respect to the intrafamily control group, revealing a significant gap, which becomes more evident at very old ages.

In order to quantify the survival advantage due to a presence of a long-lived subject in the family, the siblings' hazard function was compared with those of their spouses by means of a Cox regression model. In this model, 'relationship to the proband', 'gender of the sibling/spouse' and their interaction were used as explanatory covariates. In Table 3, the maximum likelihood estimation of the parameters of this model and the hazard ratio (HR) for mortality risk are reported. From this model, a significant survival advantage for male siblings of probands is shown. In fact, they have a substantial mortality reduction of about 28% ($e^{-0.005-0.325}$) when compared with the spouses of female siblings (HR=0.719). Also adjusting for cohort effect (by inserting the year of birth of siblings/spouses as adjunctive covariate in the model) this reduction remained almost constant (data not shown).

In order to further investigate whether the sex of the proband had an effect on the survival probabilities of their siblings, we split the data set according to the sex of the proband. In 76 out of 202 families the sex of the proband was male. Figures 1 and 2 of the Supplementary Material show the survival curves of the siblings of probands and those of their spouses according to the sex of the proband. When the sex of the proband was male (Supplementary Figure 1), both male and female siblings had a survival advantage with respect to their spouses ($P=0.029$ for males; $P=0.037$ for females). When only families with a female proband were analyzed (Supplementary Figure 2), only male siblings showed a survival advantage with respect to the intrafamily control group ($P=0.007$). Parallel results were obtained by Cox regression analysis (see Table 4). In fact, siblings of male probands had a mortality reduction of about 23% with respect to their spouses (HR=0.772; $P=0.004$). On the contrary, when the sex of the proband was female, only male siblings showed such a survival advantage. In fact the interaction term of the correspondent model indicated that male siblings had a significant mortality reduction of about 30% ($e^{0.144-0.494}$).

It is of note that, although life tables show that women live longer than males (about 5 years), none of the results obtained here differed if we considered different cut offs (between 91 and 99) to define female probands.

DISCUSSION

For years, the reduced mortality of family members of centenarians has suggested the presence of a genetic component in the longevity trait. However it has always been very clear to scientists studying this issue that environmental and familiar factors (such as economic and social status) could influence the probability of attaining longevity together with genetics. In addition, it is well known that the heritability of a trait is population specific, as it may be influenced by different factors acting differently on certain traits in different populations. This is probably particularly true for longevity, which is increasing due to environmental factors (better food, better medical assistance and so on) across western countries but at different speeds. It is then likely that the importance of genetics on longevity may be higher in areas with slower yet more recent progress (such as Calabria and Sardinia) than in other areas of Western Countries.[17,18] Finally, many cues support the hypothesis that the heritability of longevity might be higher in males than in females.

The present study has confirmed the presence of a strong familiar component on longevity. In fact both the parents and the siblings (either females and males) of long-lived probands were found to live longer than the general contemporary population. On the other hand, the comparison of survival curves of the siblings of nonagenarians with those of their spouses (which are genetically unrelated but share a great part of their environment) shows a slightly different picture. In fact, we found that brothers of nonagenarians lived significantly longer than the husbands of their sisters. By contrast, no difference could be detected between survival curves of sisters of nonagenarians and the survival curves of the wives of their brothers, suggesting that the heritability of longevity is higher in males than in females. This is further reinforced by the subsequent observation that the siblings of

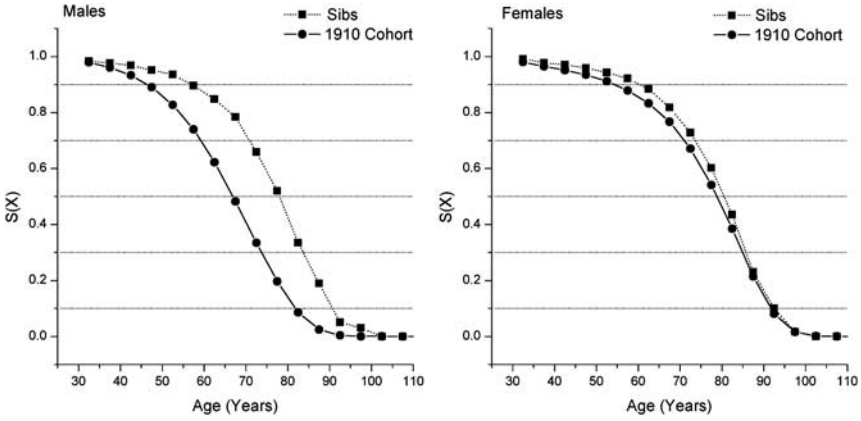


Figure 1 Survival probabilities from age 30 for siblings of probands with respect to the Italian 1910 cohort by gender.

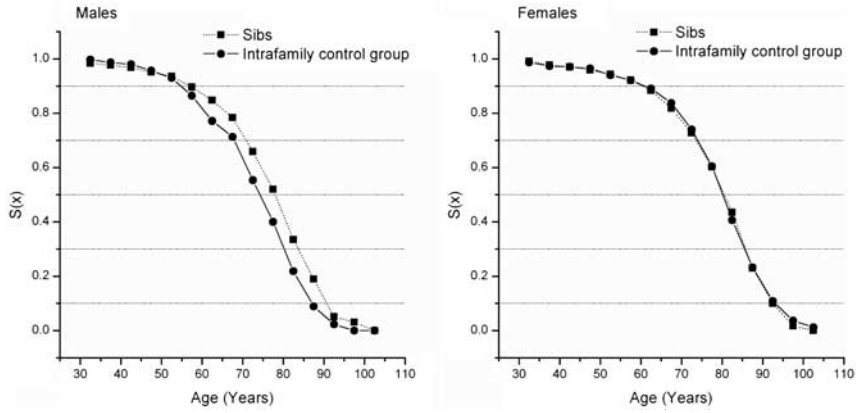


Figure 2 Survival probabilities from age 30 for siblings of probands with respect to the relevant intrafamily control group.

Table 3 Maximum likelihood estimation of the parameters of the fitted Cox proportional hazards model

						95% CI for HR	
Variables	*Coefficient (β)*	*SE*	*Wald*	*P-value**	*HR*	*Lower*	*Upper*
Relation to the proband=sibling	−0.005	0.081	0.003	0.953	0.995	0.849	1.167
Gender of the sibling/spouse=female	0.521	0.087	36.260	0.000	1.684	1.422	1.996
Relation to the proband*, gender of the sibling/spouse	−0.325	0.114	8.182	0.004	0.723	0.578	0.903

Abbreviations: CI, confidence interval; HR, hazard ratio.
SE of the estimated coefficients with the relevant HR and CI of the model are reported.
**P*-values refer to the Wald tests.

male probands (either males or females) show a reduced mortality than their spouses. By contrast, when we analyzed the siblings of female probands we found that only their brothers had a lower mortality when compared with the male spouses. This result suggests that, independently of gender, family members with a male proband share, on average, a significant genetic advantage. On the other hand, in the sibships with a female proband, the genetic share of the familial advantage is on average lower, and the female spouses of brothers of nonagenarians benefit most from the familial advantage. These results confirm that longevity has a genetic component, and suggest that such a component is stronger in males than in females. On the other hand, they also suggest that females can take advantage of a favorable environment more than males. In fact, we may state that, according to our data, being the sister of a long-lived subject or marrying one of the brothers of this subject provides a woman almost with the same survival advantage.

It is certainly important to outline some limitations of the study. First of all it is important to point out that our results may be in part specific to a largely rural and underdeveloped society where social differences are very strong, especially until a few decades ago.[23] In fact, in contrast to the study of Shoenmaker *et al*,[3] spouses of proband's siblings also live longer than the corresponding birth cohort. It is also worth mentioning that males in these cohorts may have taken advantages of their families more than their sisters in terms of wealth and social benefits. Indeed, we previously showed that only a very small percentage of women born around the beginning of the XX

Table 4 Maximum likelihood estimation of the parameters for the Cox regression models with respect to the sex of the proband

Variables	*Coefficient (β)*	*SE*	*Wald*	P-*value**	*HR*	*95% CI for HR*
(a) Sex of the proband=female						
Relation to the proband=sibling	0.144	0.104	1.902	0.168	1.155	0.941–1.417
Gender of the sibling/spouse=female	0.691	0.113	37.178	<0.001	1.996	1.598–2.492
Relation to the proband *, gender of the sibling/spouse	−0.494	0.147	11.305	0.001	0.610	0.458–0.814
(b) Sex of the proband=male[a]						
Relation to the proband=sibling	−0.258	0.090	8.165	0.004	0.772	0.647–0.922
Gender of the sibling/spouse=female	0.254	0.090	8.035	0.005	1.289	1.082–1.537

Abbreviations: CI, confidence interval; HR, hazard ratio.
SE of the estimated coefficients with the relevant HR and CI of the model are reported.
[a]The interaction term was not significant (β=−0.092; P=0.607; HR=0.912 with a CI=0.641–1.297).
**P*-values refer to the Wald tests.

century were properly scholarized.[23] This may partly explain the small excess in survival of sisters over the wives of the brothers or over the birth cohort as compared with the same groups in men.

In addition we need to point out that we used life tables referring to the 1910 Italian birth cohort as for that period they are not available for the Calabrian population alone. Calabrian life tables from 1940s onward do not show significant differences with respect to the average Italian mortality data. However, we may suppose that, based on its socio economic conditions,[24] life expectancy in Calabria at the beginning of the XX century was lower than in the rest of Italy, where, on turn, it was lower than in northern European countries.[25] Therefore we can expect that this point does not affect our results or led to an underestimate of survival advantage with respect to the general population cohorts. On the other hand, our results are in agreement with numerous demographic reports showing that in the last decades, where medical and social conditions have greatly improved, the increase in the number of female centenarians in Europe has been by far faster than the increase of male centenarians.[26]

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The study was supported by Fondi di ateneo (ex 60%). The support of the municipalities (Bisignano, Cariati, Cosenza, Luzzi, Montalto Uffugo, Rende, and Rose) and of their officers is warmly acknowledged.

1 Perls TT, Bubrick E, Wager CG, Vijg J, Kruglyak L: Siblings of centenarians live longer. *Lancet* 1998; **351**: 1560.
2 Atzmon G, Rincon M, Rabizadeh P, Barzilai N: Biological evidence for inheritance of exceptional longevity. *Mech Ageing Dev* 2005; **126**: 341–345.
3 Schoenmaker M, de Craen AJ, de Meijer PH *et al*: Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur J Hum Genet* 2006; **14**: 79–84.
4 Perls TT, Wilmoth J, Levenson R *et al*: Life-long sustained mortality advantage of siblings of centenarians. *Proc Natl Acad Sci USA* 2002; **99**: 8442–8447.
5 Gudmundsson H, Gudbjartsson DF, Frigge M, Gulcher JR, Stefansson K: Inheritance of human longevity in Iceland. *Eur J Hum Genet* 2000; **8**: 743–749.
6 Kerber RA, O'Brien E, Smith KR, Cawthon RM: Familial excess longevity in Utah genealogies. *J Gerontol A Biol Sci Med Sci* 2001; **56**: B130–B139.
7 Terry DF, Wilcox M, McCormick MA, Lawler E, Perls TT: Cardiovascular advantages among the offspring of centenarians. *J Gerontol A Biol Sci Med Sci* 2003; **58**: M425–M431.
8 Terry DF, Wilcox MA, McCormick MA *et al*: Lower all-cause, cardiovascular, and cancer mortality in centenarians' offspring. *J Am Geriatr Soc* 2004; **52**: 2074–2076.
9 Willcox BJ, Willcox DC, He Q, Curb JD, Suzuki M: Siblings of Okinawan centenarians share lifelong mortality advantages. *J Gerontol A Biol Sci Med Sci* 2006; **61**: 345–354.
10 Atzmon G, Schechter C, Greiner W, Davidson D, Rennert G, Barzilai N: Clinical phenotype of families with longevity. *J Am Geriatr Soc* 2004; **52**: 274–277.
11 Cournil A, Legay JM, Schachter F: Evidence of sex-linked effects on the inheritance of human longevity: a population-based study in the Valserine valley (French Jura), 18–20th centuries. *Proc Biol Sci* 2000; **267**: 1021–1025.
12 Herskind AM, McGue M, Holm NV, Sorensen TI, Harvald B, Vaupel JW: The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. *Hum Genet* 1996; **97**: 319–323.
13 Ljungquist B, Berg S, Lanke J, McClearn GE, Pedersen NL: The effect of genetic factors for longevity: a comparison of identical and fraternal twins in the Swedish Twin Registry. *J Gerontol A Biol Sci Med Sci* 1998; **53**: M441–M446.
14 Vaupel JW, Carey JR, Christensen K *et al*: Biodemographic trajectories of longevity. *Science* 1998; **280**: 855–860.
15 Kirkwood TB: Time of our lives. What controls the length of life? *EMBO Rep* 2005; **6**: Spec no: S4–S8.
16 Yashin AI, Ukraintseva SV, De Benedictis G *et al*: Have the oldest old adults ever been frail in the past? A hypothesis that explains modern trends in survival. *J Gerontol A Biol Sci Med Sci* 2001; **56**: B432–B442.
17 Montesanto A, Passarino G, Senatore A, Carotenuto L, De Benedictis G: Spatial analysis and surname analysis: complementary tools for shedding light on human longevity patterns. *Ann Hum Genet* 2008; **72**: 253–260.
18 Poulain M, Pes GM, Grasland C *et al*: Identification of a geographic area characterized by extreme longevity in the Sardinia island: the AKEA study. *Exp Gerontol* 2004; **39**: 1423–1429.
19 Perls T, Kohler IV, Andersen S *et al*: Survival of parents and siblings of supercentenarians. *J Gerontol A Biol Sci Med Sci* 2007; **62**: 1028–1034.
20 Brillinger DR: The natural variability of vital rates and associated statistics. *Biometrics* 1986; **42**: 693–734.
21 Elandt-Johnson RC, Johnson NL: *Survival models and data analysis*. Wiley: New York, 1980.
22 Cox DR: Regression Models and Life-Tables. *J R Stat Soc Series B (Methodol)* 1972; **34**: 187–220.
23 De Rango F, Montesanto A, Berardelli M *et al*: To grow old in southern Italy: a comprehensive description of the old and oldest old in Calabria. *Gerontology* 2010, in press.
24 Tagarelli A, Piro A, Tagarelli G, Zinno F: Color-blindness in Calabria (Southern Italy): a north-south decreasing trend. *Am J Hum Biol* 2000; **12**: 17–24.
25 Jeune B, Skytthe A, Cournil A *et al*: Handgrip strength among nonagenarians and centenarians in three European regions. *J Gerontol A Biol Sci Med Sci* 2006; **61**: 707–712.
26 Robine JM, Paccaud F: Nonagenarians and centenarians in Switzerland, 1860–2001: a demographic analysis. *J Epidemiol Community Health* 2005; **59**: 31–37.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)