Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2008 September ; 17(9): 2208–2214. doi: 10.1158/1055-9965.EPI-08-0183.

Family-Based Samples Can Play an Important Role in Genetic Association Studies

Ethan M. Lange[1,2], **Jielin Sun**[3,4], **Leslie A. Lange**[1], **S. Lilly Zheng**[3,4], **David Duggan**[5], **John D. Carpten**[5], **Henrik Gronberg**[6], **William B. Isaacs**[7], **Jianfeng Xu**[3,4,†], and **Bao-Li Chang**[3,4]

1*Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC*

2*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC*

3*Center for Cancer Genomics, Wake Forest University School of Medicine, Winston-Salem, NC*

4*Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, NC*

5*Translational Genomics Research Institute (TGen), Phoenix, AZ*

6*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

7*Johns Hopkins Medical Institutions, Baltimore, MD*

Abstract

Over the past two decades, DNA samples from thousands of families have been collected and genotyped for linkage studies of common complex diseases such as type 2 diabetes, asthma and prostate cancer. Unfortunately, little success has been achieved in identifying genetic susceptibility risk factors through these considerable efforts. However, significant success in identifying common disease risk-associated variants has been recently achieved from genome-wide association (GWA) studies using unrelated case-control samples. These GWA studies are typically performed using population-based cases and controls that are ascertained irrespective of their family history for the disease of interest. Few genetic association studies have taken full advantage of the considerable resources that are available from the linkage-based family collections despite evidence showing cases that have a positive family history of disease are more likely to carry common genetic variants associated with disease susceptibility. Herein, we argue that population stratification is still a concern in case-control genetic association studies, despite the development of analytic methods designed to account for this source of confounding, for a subset of SNPs in the genome, most notably those SNPs in regions involved with natural selection. We note that current analytic approaches designed to address the issue of population stratification in case-control studies cannot definitively distinguish between true and false associations and we argue that family-based samples can still serve an invaluable role in following-up findings from case-control studies.

Keywords

population stratification; prostate cancer; association; 8q24

†Address for correspondence: Dr. Jianfeng Xu, M.D., Dr.PH., Center for Human Genomics, Medical Center Blvd, Winston-Salem, NC 27157, Phone: (336) 713-7500, Fax: (336) 713-7566, Email: jxu@wfubmc.edu,
The authors had full responsibility for the analysis and interpretation of the data, the writing of the manuscript, and the decision to submit the manuscript for publication.

Introduction

Recent successes in identifying disease risk associated variants using genome-wide association (GWA) studies have demonstrated the power of this approach in discovering novel risk factors. Interestingly, many of these genetic risk variants were identified in study populations consisting of patients with specific diseases and a common set of controls from the general population (1), rather than in well-matched case-control populations where cases and controls are ascertained from the same well defined populations. The fact that many of these associations can be consistently replicated in other study populations appears to relieve concern about population stratification, a phenomenon in genetic association studies where differences in allele or genotype frequencies at a particular polymorphism between samples of cases and controls are due to differences in ethnic origins between the two samples, rather than a real effect of the variation at the polymorphism on disease risk. However, by carefully examining genetic association findings of several recent studies, we found that population stratification may manifest in different forms and still be a concern in genetic association studies for a subset of SNPs in the genome. We believe that proper recognition of the potential problem, utilization of appropriate study designs and analytical tools to address this potential confounding, and cautious interpretation of association results are still critical issues in genetic association studies. We argue that family-based association studies, which take advantage of previously collected family samples, can play an important role in assessing the relationship between genetic variants and disease and may be particularly valuable for following up results from case-control studies.

Large geographic variation in allele frequencies for a subset of SNPs in the genome

Considerably heterogeneous allele frequencies are observed among people from different geographic regions for a subset of SNPs in the genome. As demonstrated in the Wellcome Trust Case Control Consortium (WTCCC) GWA study where allele frequencies of SNPs across the genome were compared among ~17,000 subjects from 12 broad geographical regions across Great Britain (1), SNPs in several genomic regions had highly significant differences in allele frequencies (Fig 1A), including a number of regions not previously implicated by past studies. Some of the differences reached genome-wide significance levels ($P < 10^{-7}$), for example SNPs at 2q21.3, 4p14, 6p21. Spurious statistical differences in allele frequencies between cases and controls can occur for SNPs at these regions if they are not well matched in terms of ancestry. While the impact of ancestral variation on genetic association studies can be minimized with appropriate study designs and analytical adjustment, it may be difficult to remove this confounder in all situations, such as when there are strong differences in allele frequencies among SNPs in genomic regions under selective pressure between geographically neighboring populations or when the same factors play a role in contributing to both geographical variation in allele frequencies and disease risk.

For example, allele frequencies of multiple SNPs at 4p14 near the *TLR-6-1-10* gene cluster were found to be significantly different among subjects from the 12 broad geographical regions across Great Britain in the WTCCC study; *P* values of $\sim 10^{-40}$ were found when using 11-d.f. tests (1). It is hypothesized that these differences may reflect the outcome of natural selection due to the roles of these genes in preventing a variety of infectious diseases (2). In the meantime, sequence variants in the *TLR6-1-10* gene cluster have been reported to be associated with prostate cancer risk in a well matched case-control study population from Sweden by our group (3). Several SNPs were significantly associated with prostate cancer risk in a population-based case-control study from Sweden that included 2,887 cases and 1,715 controls. For example, we found a SNP at 5′UTR of TLR1 gene (rs5743551) was significantly associated with prostate cancer risk; the minor allele frequency was 0.254 in cases and 0.221 in controls, $P = 0.0009$

(Fig 1B). Further examination of this SNP revealed that the significant association was primarily driven by subjects from the central part of Sweden, $P = 0.002$. The association was in the same direction but was not statistically significant among subjects from the northern part of Sweden. The difference in strength of association for this SNP in these two regions could be partially explained by the large observed geographic variation in minor allele frequency. The difference in the minor allele frequency observed in these two regional populations was significantly different both before and after accounting for prostate cancer ($P < 0.001$ and $P = 0.01$, respectively). Although the SNP remained significant after adjusted for geographic region ($P = 0.004$), we cannot completely remove the potential confounding effect of geographic variation because there may be additional hidden population structure within these regions. Considering that *TLR6-1-10* sequence variants may play a role in both prostate cancer risk (4,5) and natural selection (2), it is difficult to dissect true prostate cancer association from confounding using frequency tests, even in this relatively homogeneous and well matched Swedish study population.

Variable ancestry proportions in cases and controls within a race group

It is well known that false positive associations may occur in genetic association studies if samples of cases and controls are from different populations and if the rate of disease is different in the two populations. The problem likely also exists when cases and controls are ostensibly from the same race and ethnic group but have different proportions of sub-race or sub-ethnic ancestry. For example, African Americans from different parts of the U.S. have various levels of west African, east African, and European ancestry, while European Americans in the U.S. have various levels of southern or northern European ancestry (6). The magnitude of this potential problem depends on the complexity and degree of admixture among sub-groups, differences in disease risk and prevalence between the sub-groups, and differences in allele frequencies for SNPs in the genome (global) and in a specific region of genome (local) between these sub-groups. Recent results of association studies of multiple independent sequence variants at 8q24 and prostate cancer risk exemplify the complexity of this potential problem.

Association of prostate cancer risk with 8q24 variants was initially discovered in a fine mapping study of a prostate cancer linkage region at 8q24 among two case-control study populations from Iceland (7). It includes a SNP, rs1447295, and multiple other variants that are in strong linkage disequilibrium (LD) with this SNP. This locus was independently confirmed in all published study populations, including those from Sweden and the U.S (7), Multiethnic cohort (MEC) (8), Australia (9), Mayo Clinics (10), Breast and Prostate Cancer Cohort Consortium (11), and King County of Washington (12). This locus was also implicated as one of the strongest prostate cancer associated regions in three GWA studies (13-16). Interestingly, in addition to the locus rs1447295 at 128,554,220 bp (referred to as Region 1 of 8q24), two additional loci that are within 500 Kb proximal to Region 1 were found to be independently associated with prostate cancer risk (13,14,17). One of these two regions includes the SNP rs6983267 at 128,510,352 bp (referred to as Region 3 of 8q24) and the other includes a SNP rs16901979 at 128,194,098 bp (referred to as Region 2 of 8q24). Independent support for association of these three regions at 8q24 with prostate cancer risk was also observed in a hospital based case-control study from the Johns Hopkins Hospital (18). While these consistent findings strongly suggest the presence of true disease causal variants in this region, several interesting observations suggest that we need to take residual population stratification into consideration in order to appropriately interpret the observed associations at multiple independent regions of 8q24.

One interesting observation is that all of the ‘risk alleles’ for SNPs at the three 8q24 regions have higher estimated frequencies in African study populations; allele A of rs1447295 at Region 1, allele T of rs6983267 at Region 3, and allele A of rs16901979 at Region 2 were all

found to be more common in prostate cancer cases than in controls of European Americans (13-18). All of these risk alleles have considerably higher frequencies in the African population (YRI) than the European population (CEU); their frequencies (YRI/CEU) for these three SNPs in Regions 1, 3, and 2 are 0.54/0.02, 0.98/0.46, and 0.34/0.07, respectively. The implication of these large differences in allele frequencies between YRI and CEU samples is that if prostate cancer cases in general have a higher proportion of African ancestry than controls, these SNPs would be particularly susceptible to be erroneously statistically associated with prostate cancer risk due to population stratification given the increased risk for prostate cancer in men of African descent.

Data from an admixture mapping study in African American prostate cancer cases and controls does indicate that African American prostate cancer cases have a higher proportion of African ancestry than controls. As shown in Figure 2, the proportion of African ancestry was estimated at each region of the genome among African American prostate cancer cases and controls in the U.S. (8,17). In the samples considered, African American prostate cancer cases have a higher proportion of African descent across the entire genome than do African American controls. The mean proportion of African ancestry in the genome (global) was significantly higher in these cases (78.2%) than in controls (74.8%), $P < 0.001$. The difference was largest (> 7%) at a local region of 8q24 (126-129 Mb). These results, if confirmed, may have several important implications. The systematic difference in estimated proportion of African Ancestry suggests these African American samples are not ideally matched to control the effects of population stratification. The higher African ancestry proportion across the entire genome in cases than in controls suggests that many SNPs may have different allele frequencies between cases and controls due to population stratification. Correcting for this consistently observed difference in proportion of African Ancestry across the genome using appropriate analytic techniques on available genetic marker data is necessary to maintain unbiased association tests given the systematic differences in allele frequencies between European and African populations.

Another observation is that there is considerable variation in allele frequencies for the 8q24 risk variants among European subjects. As shown in Table 1, the allele frequencies for rs1447295 at Region 1 differed considerably among different European study populations. The frequency of the risk allele A ranges from 0.07 in a French population (Southern Europe) to 0.17 in a Finnish population (Northern Europe). However, it is striking to note that prostate cancer cases have a consistently higher frequency than controls within each of these study populations. One of the most favorable interpretations of these results is that the consistent association finding is due to true prostate cancer risk loci at 8q24. Alternatively, the observed association could also be due to systematic population stratification. If prostate cancer cases in each of these study populations have higher proportions of African ancestry than the controls, then we would expect to observe consistently higher frequencies of these risk alleles in cases than in controls. Unfortunately, there is a lack of data at this time for estimating African proportions among European prostate cancer cases and controls.

Analytic adjustment for case-control association tests

We have presented two scenarios where population stratification may confound results and compromise the interpretation of genetic association studies. As our knowledge about genetic variations among different human populations improves, we will have a better appreciation of the various forms and extent of population stratification on genetic association studies. Analytical methods and study designs that address subtle and complex population stratification are needed to address these concerns.

Currently, several methods are available to adjust for potential population stratification. These methods include genomic control (19), structured association (20,21), and principal components analysis (22). All three methods have been demonstrated to reduce the impact of population stratification in genetic association studies. In addition, all three procedures require the additional genotyping of a sizable number of genetic markers, preferably not in linkage disequilibrium, that are hypothesized not to be associated with the trait under study. Genomic control corrects for stratification by adjusting association test statistics at each marker by a uniform inflation factor determined by evaluating the distribution of test statistics across all markers. However, this approach does not factor in that some markers will have greater differences in alleles frequencies across different ancestral populations than others. Thus, the uniform adjustment would under correct for markers with large differences in allele frequencies across populations and over correct for markers with relatively small differences in allele frequencies across populations. Structured association is based on assigning individuals to subpopulations using a model-based Bayesian clustering algorithm implemented in the program STRUCTURE, and then carrying out all analyses conditional on the inferred assignments. Given the computational complexity of this approach, structured association is limited to a relatively small number of ancestral informative markers – markers that have large allele frequencies across measured populations. The primary concerns regarding this approach are the reliance on a relatively small number of markers to differentiate between populations, the adequacy of a modest number of markers to differentiate between populations that have not been previously extensively studied, and the sensitivity of the approach to the assumed number of clusters (which is somewhat arbitrarily chosen by the user). The utilization of principal components, a data reduction method designed to capture most of the variability across all SNPs using a relatively small number of independent continuous variables, allows the inclusion of a genome-wide panel of markers to measure and account for systematic differences in ancestry is becoming an increasingly popular analytic approach for identifying population substructure and for controlling for it in genetic association studies. This method has the benefit of being computationally efficient and allows the inclusion of hundreds of thousands of genetic markers, making the *a priori* choice of ancestry informative markers in GWA studies unnecessary.

Critical to the validity of case-control genetic association study results is the ability of analytic methods to account for population stratification, particularly for SNPs in regions of selection. Despite the recognition of this problem, empirical and simulation-based studies designed to evaluate the impact of population stratification, after application of methods designed to control for it, have been somewhat limited. It should be noted that all of the described analytic approaches failed to resolve the observed population stratification in a case-control sample of Americans of European origin that were selected for extreme values of height (6,22,23). For example, in the study by Campbell and colleagues (6), height and allele frequencies at the lactase gene (LCT [MIM 603202]) varied considerably from northwestern to southeastern Europe and a false-positive association was found linking the lactase gene with height in this sample. Results across a number of studies suggest that analytic methods designed to address confounding will dramatically reduce the number of false positives due to population stratification but will not completely eliminate the problem (6). A recent European GWA study on Rheumatoid arthritis found evidence for spurious association at *LCT* and *IRF4*, two regions that are known to be highly selected in European populations (24). The application of both principal components and STRUCTURE, in this study, reduced the confounding at these two regions, but did not remove the effects entirely. Regardless, if the causal polymorphisms are highly correlated with underlying population structure (e.g. because of natural selection), it will not be possible to distinguish between true and false positives statistically, and any attempt to remove the population structure will reduce the power to detect truly associated markers.

The recent WTCCC GWA study found only modest evidence of over-dispersion (λ ranging from 1.03-1.11 for the seven diseases evaluated) in their association trend test statistics, despite strong evidence for highly discrepant allele frequencies at SNPs in 13 different chromosomal regions between control samples collected from 12 different geographical regions in the UK (1). In this study, both cases and controls were ascertained throughout the UK; there was little evidence for confounding due to population stratification when comparing case allele frequencies to that of controls even prior to incorporating analytic methods to control for this effect. On the other hand, the observed evidence for numerous genomic regions under apparent selective pressure in the WTCCC study could cause some concern regarding the validity of case-control genetic association study results for a subset of SNPs in the genome, especially when cases and controls are not ascertained from the same geographical region.

Family-based samples and family-based genetic association studies

For many complex diseases, a large number of pedigrees with multiple affected relatives have been recruited in the past for genetic linkage studies. While the ability of genetic linkage studies to identify regions that contain susceptibility variants for complex diseases may be limited due to heterogeneity, reduced penetrance, and phenocopies (25), these families may still be extremely useful for association testing. Of note, DNA samples from thousands of families with hereditary prostate cancer have already been collected for linkage studies. These families have extensive clinical information with respect to prostate and other cancers in the probands and other family members that may prove invaluable when trying to understand the complex genetic etiology of the disease and offer the unique opportunity to evaluate parent-of-origin effects, or imprinting, that cannot be evaluated in studies of unrelated cases and controls.

Typical genetic case-control association studies use unrelated population-based cases and controls. Several theoretical and simulation studies (26-29) have demonstrated that greater power can be achieved in genetic case-control association studies by selecting cases from families with multiple cases over selecting unrelated cases with no family history of the disease. Empirical results from prostate cancer genetic association studies support these conclusions. Specifically, we observed stronger effects of 8q24 variants in our prostate cancer cases from high-density hereditary prostate cancer families than in unrelated prostate cancer cases collected independent of family history (30). A recently completed GWA study for prostate cancer using cases diagnosed prior to age 61 or with a family history of disease was tremendously successful in not only strongly replicating the results from previous GWA studies (namely, variants associated with prostate cancer on 8q and 17q) but also in identifying seven new regions that harbor genetic variants significantly associated with the disease (16). This study demonstrates the considerable power that can be gained by studying cases, such as those collected through linkage studies, which are enriched for carrying genetic susceptibility variants.

While considerable power can be gained using familial cases in case-control association studies, the use of familial cases in the case-control study design does not protect the conclusions from the impact of population stratification. An alternative study design to case-control designs is family-based association. Family-based association methods evaluate whether particular alleles are transmitted from parents down to affected offspring in a proportion that is different than expectation under the null hypothesis of no association between marker and disease. Because these methods utilize non-transmitted alleles from the same parents as the control sample, these methods are not susceptible to population stratification. The transmission disequilibrium test was originally proposed for a parent-parent-affected offspring design (31), but has been extended to include siblings discordant for the disease under study and general family-designs (32,33). Since many of the diseases of interest typically have

a late age of onset, the consequential unavailability of parents has made these more general forms of the transmission disequilibrium test popular.

While family-based study designs offer protection from false positives due to population stratification, they have not been the design of choice for many due to the relative inefficiency of the design. Specifically, it is widely accepted that collecting family-based samples is considerably more expensive than collecting a sample of unrelated cases and controls. In addition, the matching of transmitted with non-transmitted alleles from the same family reduces statistical power. Consequently, to get the same power as a case-control study design, a larger number of family-based samples will have to be ascertained and genotyped. This increased cost has motivated the development of statistical techniques to account for population stratification in case-control designs. However, as noted above, these statistical methods are not perfect and cannot completely remove doubt regarding the validity of any single association result in a case-control design. Many family-based samples have already been collected through linkage studies, thus the additional cost of collecting these family-based samples is often not a major concern. The additional costs associated with having to genotype a significantly larger number of samples to achieve equivalent power still likely renders performing a GWA study on family-based samples (even on previously collected samples) inefficient. Unique to nuclear families when both parents are available for genotyping, the cost of performing a GWA study can be mitigated by genotyping the offspring on a significantly reduced number of SNPs and imputing (with a high degree of accuracy) the remaining SNPs in the offspring by using the linkage phase on the subset of SNPs typed for all family members.

Family-based studies may be ideal for validating previously identified genetic risk factors, Follow-up case-control studies typically have to perform extra genotyping for a large number of ancestry informative markers (which would not need to be genotyped for a family-based study) in an attempt to control for population stratification. In addition, while GWA studies require a strict correction for multiple testing to account for the hundreds of thousands of genetic markers analyzed, follow-up studies require a much less stringent significance threshold given the focused a priori hypotheses. Combining these factors with the fact that large family-based samples have already been collected suggests family-based association tests can be an efficient and useful strategy for validating previously identified genetic susceptibility markers. Further studies need to be performed to evaluate the relative power of family-based association analyses compared to case-control designs when utilizing methods that correct for population stratification. In particular, it is important to evaluate the impact of poorly matched case-control samples, over correction for ancestry, and causal markers that are strongly associated with population structure. Regardless of these findings, family-based association methods are clearly useful to dissect true association from false association due to population stratification.

We performed a transmission disequilibrium test for SNPs at Regions 1, 2 and 3 of 8q24 in 168 HPC families of European ancestry using the FBAT program (30). As shown in Table 2, risk alleles of one SNP in Region 2 was significantly over-transmitted from parents to affected men ($P = 0.003$). However, this was not the case for risk alleles of SNPs in Regions 1 and 3, despite significant differences in allele frequencies observed for these SNPs between these same familial cases and a sample of unrelated screened controls. Results from the family-based transmission tests suggested that SNPs at Region 2 confer prostate cancer risk while creating some concern that the association between prostate cancer and SNPs in the other two regions may be due to population stratification (perhaps a consequence of higher African ancestry at 8q24 in prostate cancer cases than controls).

From a clinical standpoint, common susceptibility markers found in primarily non-familial cases needs to be evaluated in familial cases. Individuals with a strong family history of a

disease are most likely to seek genetic counseling and subsequent genotyping for genetic risk factors for that disease. While family-based samples collected for linkage analyses are a non-representative collection of cases that are more likely to carry genetic susceptibility than population based cases, these existing family data can be useful to assess how important these risk factors are in individuals most likely to seek medical prevention. Finally, for diseases such as prostate cancer, genetic heterogeneity has made the identification of rare highly penetrant mutations through linkage analysis difficult. Evaluating the role of common variants in hereditary prostate cancer families will be critical in determining whether an accumulation of common risk variants is responsible for hereditary prostate cancer or whether there exist a number of rare high penetrant mutations that explains the significant clustering of the disease within families.

Conclusion

We believe that cautious interpretation and continued attention to population stratification as a potential confounder remains a critical issue in case-control genetic association studies in spite of the recent advances in analytical methods designed to address this potential problem. Family-based association study designs are largely impervious to the effects of population stratification in terms of controlling the rate of false positive tests. Family-based samples are readily available from linkage studies that have collected DNA samples from thousands of pedigrees that have multiple cases of disease. Despite some clear advantages, including higher rates of genetic susceptibility variants in familial cases, these family samples have been largely ignored in GWA studies. Family-based association study designs have not been routinely used because of concerns that they are not as statistically powerful or cost effective as case-control designs. These concerns argue that family-based association methods are not ideal for SNP discovery. However, family-based association studies may be vital in teasing out which variants identified in case-control studies are truly associated with disease and not an artifact of population stratification.

Acknowledgements

Funding

The effort is partially supported by National Cancer Institute CA106523, CA95052, and CA105055 to J.X., CA112517 and CA58236 to W.B.I., CA86323 to A.W.P., CA79596 and CA119069 to E.L., CA119069 to B.L.C., and Department of Defense grant PC051264 to J.X.

References

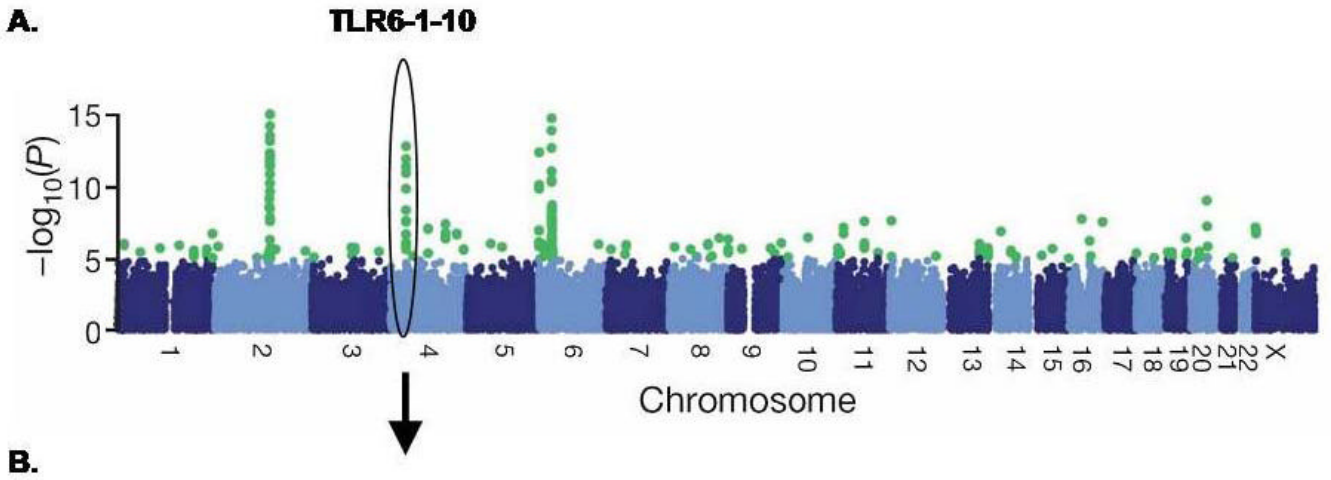
1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78. [PubMed: 17554300]
2. Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39:857–64. [PubMed: 17554260]
3. Sun J, Wiklund F, Zheng SL, et al. Sequence variants in Toll-like receptor gene cluster (TLR6-TLR1-TLR10) and prostate cancer risk. J Natl Cancer Inst 2005;97:525–32. [PubMed: 15812078]
4. Zheng SL, Liu W, Wiklund F, et al. A comprehensive association study for genes in inflammation pathway provides support for their roles in prostate cancer risk in the CAPS study. Prostate 2006;66:1556–64. [PubMed: 16921508]
5. De Marzo AM, Platz EA, Sutcliffe S, et al. Inflammation in prostate carcinogenesis. Nat Rev Cancer 2007;7:256–69. [PubMed: 17384581]Review
6. Campbell CD, Ogburn EL, Lunetta KL, et al. Demonstrating stratification in a European American population. Nat Genet 2005;37:868–72. [PubMed: 16041375]
7. Amundadottir LT, Sulem P, Gudmundsson J, et al. A common variant associated with prostate cancer in European and African populations. Nat Genet 2006;38:652–8. [PubMed: 16682969]

8. Freedman ML, Haiman CA, Patterson N, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 2006;103:14068–73. [PubMed: 16945910]
9. Severi G, Hayes VM, Padilla EJ, et al. The common variant rs1447295 on chromosome 8q24 and prostate cancer risk: results from an Australian population-based case-control study. Cancer Epidemiol Biomarkers Prev 2007;16:610–2. [PubMed: 17372260]
10. Wang L, McDonnell SK, Slusser JP, et al. Two common chromosome 8q24 variants are associated with increased risk for prostate cancer. Cancer Res 2007;67:2944–50. [PubMed: 17409399]
11. Schumacher FR, Feigelson HS, Cox DG, et al. A common 8q24 variant in prostate and breast cancer from a large nested case-control study. Cancer Res 2007;67:2951–6. [PubMed: 17409400]
12. Suuriniemi M, Agalliu I, Schaid DJ, et al. Confirmation of a positive association between prostate cancer risk and a locus at chromosome 8q24. Cancer Epidemiol Biomarkers Prev 2007;16:809–14. [PubMed: 17416775]
13. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645–9. [PubMed: 17401363]
14. Gudmundsson J, Sulem P, Manolescu A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007;39:631–7. [PubMed: 17401366]
15. Thomas G, Jacobs KB, Yeager M, et al. Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 2008;40:310–315. [PubMed: 18264096]
16. Eeles RA, Kote-Jarai Z, Giles GG, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 2008;40:316–321. [PubMed: 18264097]
17. Haiman CA, Patterson N, Freedman ML, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 2007;39:638–44. [PubMed: 17401364]
18. Zheng SL, Sun J, Cheng Y, et al. Association between two unlinked loci at 8q24 and prostate cancer risk among European Americans. J Natl Cancer Inst 2007;99:1525–33. [PubMed: 17925536]
19. Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;55:997–1004. [PubMed: 11315092]
20. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220–8. [PubMed: 10364535]
21. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–59. [PubMed: 10835412]
22. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–9. [PubMed: 16862161]
23. Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 2007;80(5):921–30. [PubMed: 17436246]
24. Tian C, Plenge RM, Ransom M, et al. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet Jan;2008 4(1):e4. [PubMed: 18208329]
25. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–7. [PubMed: 8801636]
26. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 1998;8:1273–88. [PubMed: 9872982]
27. Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000;405:847–56. [PubMed: 10866211]
28. Fingerlin TE, Boehnke M, Abecasis GR. Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. Am J Hum Genet 2004;74:432–43. [PubMed: 14752704]
29. Li M, Boehnke M, Abecasis GR. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet 2006;78:778–92. [PubMed: 16642434]
30. Sun J, Lange EM, Isaacs SD, et al. Chromosome 8q24 risk variants in hereditary and non-hereditary prostate cancer patients. Prostate 2008;68:489–97. [PubMed: 18213635]

31. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulindependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516. [PubMed: 8447318]

32. Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 1998;62:450–458. [PubMed: 9463321]

33. Lake SL, Blacker D, Laird NM. Family-based tests of association in the presence of linkage. Am J Hum Genet 2000;67:1515–1525. [PubMed: 11058432]



Different allele frequencies of SNP rs5743551 in the *TLR1* gene at 4p14 among prostate cancer cases and controls in Sweden

	Number of subjects			Allele freqquencies		
	Cases	Controls	Allele	Cases	Controls	*P-dom*
CAPS	2,887	1,715	C	0.254	0.221	0.0009
Region 1 (Central)	1,059	394	C	0.273	0.259	0.6
Region 2 (Northern)	1,828	1,321	C	0.244	0.209	0.002

Figure 1.

Geographic variation of allele frequencies in the genome. A: Copy of the Figure 2A in the paper (1) showing P values for the 11-d.f. test for differences in SNP allele frequencies between 12 broad geographical regions in Great Britain. Green dots indicate SNPs with a P value < 1×10^{-5}. Cluster of highly significant SNPs are found in several genomic regions, including 4p14 that contains *TLR6-1-10* gene cluster. B: Association test between prostate cancer risk and a SNP at 5′UTR of *TLR1* gene (rs5743551) in the CAPS (Cancer of Prostate in Sweden) study population. Statistical association was found in CAPS ($P = 0.0009$). However, significant difference in allele frequencies between Region 1 (Northern part of Sweden) and Region 2 (Central part of Sweden) were also found, although the minor allele was always higher in cases than in controls in both regions.

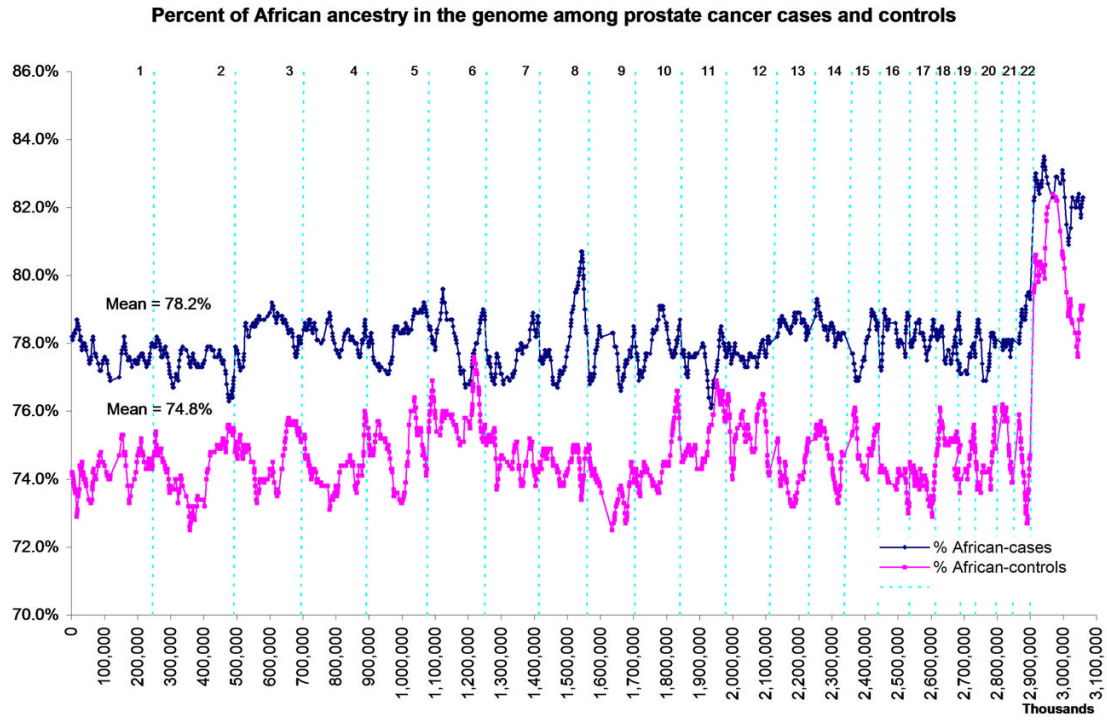


Figure 2.

Estimated percent of African ancestry in the genome among prostate cancer cases and controls in African Americans. Plot was made based on Supplementary Table 6 of published paper (17). Higher percent of African ancestry in cases than controls was found across the genome.

Table 1

Reported frequency of allele A at rs1447295 in men of European ancestry

	# of subjects		Freq		
Study (reference)	**Cases**	**Controls**	**Cases**	**Controls**	**Note**
DeCODE (7)	1,291	997	0.17	0.11	Iceland
CAPS (7)	1,435	779	0.16	0.13	Sweden
Chicago (7)	458	247	0.13	0.08	US
Australia (9)	821	732	0.15	0.11	Australia
Mayo (10)	435	545	0.12	0.10	US
ACS (11)	1,150	1,151	0.12	0.08	US
ATBC (11,13)	894	896	0.21	0.17	Finland
EPIC (11)	732	1,114	0.13	0.12	Europe
HPFS (11,13)	625	636	0.13	0.09	US
MEC (11,17)	1,168	938	0.13	0.09	US
PHS (11)	969	1,264	0.12	0.09	US
PLCO (11,13)	1,172	1,157	0.14	0.10	US
Washington (12)	630	564	0.14	0.11	US
FPCC (13)	455	459	0.12	0.07	France
Baltimore (18)	1,545	576	0.13	0.08	US

Note: deCODE: deCODE Genetics; CAPS: Cancer of Prostate in Sweden; ACS: American Cancer Society; ATBC: Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study; EPIC: European Prospective Investigation into Cancer and Nutrition Cohort; HPFS: Health Professionals Follow-up Study; MEC: Multiethnic Cohort Study; PHS: Physicians' Health Study; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; FPCC: French Prostate Case-Control Study

Table 2

Results from family-based association tests

SNPs	Position[§]	Region	Allele	Freq	# of family	S	E(S)	Var(S)	Z	P
rs16901979	128,194,098	2	A	0.07	20	37	28.00	9.34	2.95	0.0032
rs6983267	128,482,487	3	G	0.50	70	262	258.86	47.53	0.46	0.65
rs1447295	128,554,220	1	A	0.12	47	89	84.00	25.09	1.00	0.32

[§]Position is based on NCBI Build 35.