Supplementary Note 1

Correlation of 8q24 to four subphenotypes

We explored the association of the chromosome 8q24 locus to 4 phenotypes (Table A): age of diagnosis, grade of tumor, stage of disease, and family history in a first degree relative. To test for association to each phenotype, we rank-ordered all individuals for whom the phenotype was available, and calculated a cumulative LOD score for cases with that value or less. If there is no relationship between the 8q24 locus and the subphenotype, we expect the LOD score to rise linearly toward the value when all cases are studied together (dotted line in Figure 2 of the main paper, giving the correlation to age of diagnosis), with fluctuations from this expectation due to stochastic variation in sample ordering. In practice for the age of diagnosis phenotype, we see a dramatic increase of the cumulative LOD score above expectation.

Table A: Four subphenotypes for which we explored evidence for association

Subphenotype	**Categories**
Age of diagnosis (22 categories)	Equally spaced age cutoffs from 39 to 88
Grade (4 categories)	Gleason 2-4, 5-7, 8-10, undifferentiated/anaplastic
Stage (6 categories)	(1) Local, (2) Regional by extension, (3) Regional by nodes, (4) Regional by extension & nodes, (5) Regional NOS, (6) Metastatic
Family history (2 categories)	(1) First degree relative with reported prostate cancer, or (2) not

We used a permutation analysis to formally test whether the rise or fall of the cumulative LOD score compared with expectation is significant. For each subphenotype, we carried out 1,000,000 replications in which we randomly permuted the values of the subphenotypes across all individuals, looking to see if rises or falls compared to expectation (for any value of the subphenotype) were as extreme as in the real data.

As an example for the age of diagnosis (Figure 2 in the paper), the greatest rise in the real data is 5.40 (for age < 72), and the greatest fall is -0.10 (for age < 85). Randomly permuting the ages of diagnosis 1,000,000 times, we obtained 318 examples where the score rose as high above expectation as we observed (P<0.00032 for correlation to early diagnosis), and 996,505 where the score fell as much below expectation (P<0.997 for late diagnosis). The full set of results is given in Table B.

Table B: Associations tests for each phenotype from 1,000,000 random permutations

Phenotype	**Cases with phenotype**	**Max. rise above expectation**	**P-value**	**Max. fall below expectation**	**P-value**
Age of diagnosis	1,588	5.40	*0.00032	-0.10	0.997
Grade	1,518	1.17	0.10	-0.05	0.90
Stage	1,390	0.11	0.60	-0.88	0.24
Family history	1,597	0.00	0.43	-1.04	0.10

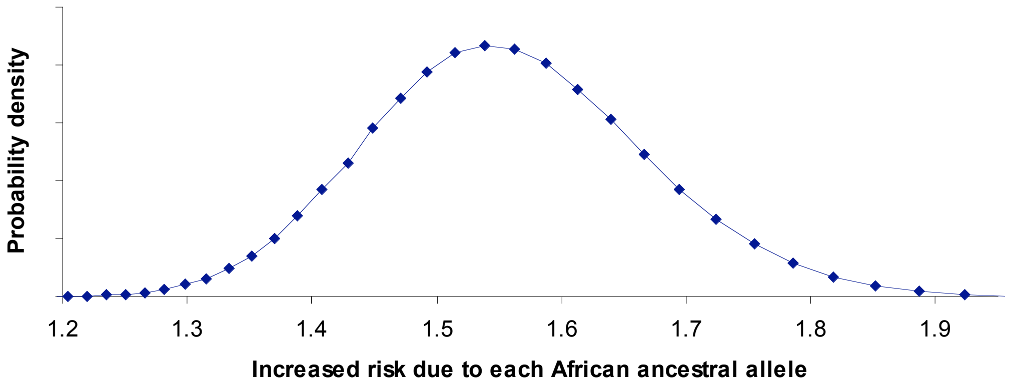
* statistically significant result

We note that Amundadottir et al. (Nature Genetics, 2006) claimed a weakly significant (P<0.02) association of the -8 allele microsatellite DG8S737 at the chromosome 8q24 locus not only to prostate cancer, but also to high grade tumor (Gleason ≥8). We do not replicate this result, and the data point in the opposite direction, indicating a suggestive association (P<0.10) to low grade tumor (Gleason <8).

Supplementary Note 2

Contribution of 8q24 to risk in African Americans

We focus on the admixture scan for cases with age of diagnosis <72 (the model giving the best evidence for association), and a range of different models of multiplicative increased risk due to each African chromosome. The posterior probability density is.



We can use this posterior probability density to calculate a best estimate (1.54) for the increased risk due to each African allele, as well as a 90% credible interval (1.38-1.75, defined as the 90% central area in this distribution). These results indicate that each copy of an African chromosome increases risk by about 54%.

To calculate the effect on disease risk in African Americans, we compare the risk for prostate cancer averaged over all African Americans in the population, to what would be expected if all African Americans had entirely European ancestry at the locus.

Table 1: Increased risk in African Americans due to 8q24 locus

# African alleles	Frequency of genotype in population (estimated from controls, who have 25.3% European ancestry overall)	Increased risk compared to all European ancestry	Weighted risk (product of two columns on left)
0	6.4%	1	0.06
1	37.8%	1.54	0.58
2	55.8%	2.37	1.32
	Total risk increased risk:		1.97

Using the 90% credible interval for increased risk due to African ancestry (1.38-1.75), the same calculations produce a range of 1.64-2.42 for the increased risk attributable to ancestry at the locus. Thus, if all African Americans had European ancestry at the locus, 1-1/1.64 = 39% to 1-1/2.42 = 59% of prostate cases would be eliminated.

Supplementary Note 3

Testing for associated alleles in African Americans, controlling for the admixture association

Overview of approach

We extended our software for admixture mapping (ref. [1]; http://genepath.med.harvard.edu/~reich) to test whether a particular allele or haplotype contributes to disease more than can be accounted for by the admixture signal.

The intuition behind this test is that if an allele is causal for prostate cancer—or in strong linkage disequilibrium with a causal allele—then it should be significantly more correlated to risk, than African ancestry state at the locus. Thus, we search for alleles that are more differentiated in frequency between cases and controls at the locus, than would be expected just from the rise in African ancestry in cases vs. controls.

The specific hypotheses we explore are whether the -8 allele of DG8S737, the A allele of rs1447295, and the haplotype combining them, are positively associated with prostate cancer in African Americans, as was hypothesized to be the case by Amundadottir et al. [2]. In particular, they found that the -8 allele confers a significant association in African Americans, with a P-value of <0.0022, and an odds ratio of ~1.60.

We were concerned that the association of the -8 allele at DG8S737 that Amundadottir et al. [2] detected might be an artifact of population stratification and the admixture association: systematic differences in ancestry between cases and controls across 8q24. To control for the possibility of population stratification between cases and controls contributing to their signal, Amundadottir et al. [2] tested for mismatching of cases and controls in their overall proportion of ancestry, and found no evidence for it. However, they did not control for local ancestry: a rise in African ancestry throughout 8q24 in cases but not controls. An admixture association would be expected to contribute to false-positive association at any allele (like -8 allele at DG8S737) that just happens to be higher in frequency in African Americans.

Details of the procedure for calculating P-values and odds ratios

We consider 3 models of association, and implement formal tests to distinguish them.

(A) The tested allele is not causal for the disease, but only in a locus with an admixture peak. In this case, the increase in allele frequency in cases should be computable simply from the difference in ancestry between cases and controls, combined with the known African and European allele frequency. A single parameter γ—the multiplicative increased risk due to carrying a European allele—is therefore used to model the risk.

(B) The allele being tested for association is causal for disease. In this case, we expect the allele to be more increased in frequency in cases vs. controls, than would be expected simply from the elevation of African ancestry proportion in cases at 8q24. We add a second parameter R_A to indicate the multiplicative increased risk due to the allele.

(C) The allele being tested for association is not itself causal, but only in linkage disequilibrium (LD) with the causal allele. In this case, the LD with the true causal allele and thus strength of association may be different when it is carried on African vs. European chromosome background. This model has three parameters, γ, R_A and R_E, to allow different risks associated with the allele on African and European chromosomes.

The tests for association are designed to distinguish whether the more complex models (B and C), fit better than the model with only an admixture association (A). Exploratory analysis (NP not shown) indicates that this has power to detect true allelic associations.

To obtain a P-value, we use a formal likelihood ratio test.

For each model (A, B and C), we proceed by finding the maximum likelihood combination of parameters, which we do by exploring a grid of parameter combinations. We then assess whether the increase in the likelihood of the data, as one moves to models with increasing numbers of parameters, provides convincing evidence of a better fit.

For each model (A,B,C) we maximize the log likelihood obtaining L_A, L_B, L_C (natural logarithms). We then form statistics $S_1 = 2(L_B–L_A)$ and $S_2 =2(L_C - L_B)$. Standard theory for likelihood ratio tests [1] shows that under the hypothesis that A is true, S_1 is asymptotically (for large data set such as we are studying) distributed as χ^2 with 1 degree of freedom, and similarly for S_2 if B is true.

None of the alleles or haplotypes previously associated to disease by Amundadottir et al. [2] gave significant evidence for association. Neither the -8 allele at DG8S737, nor the A allele at rs1447295, nor the haplotype combining both the -8 and A alleles (phased using an expectation maximization algorithm; NP unpublished), produced significant association by either the S_1 [Table 3] or S_2 tests.

To obtain an odds ratio, we use the maximum likelihood estimate assuming that the risk due to the allele is the same in chromosomes of African and European origin (R_A=R_E). A 95% Bayesian credible interval is obtained by assuming a flat prior distribution across an equally spaced mesh of γ and log(R_A) values. Further details of the methods are in preparation for publication (NP, AT and DR).

References:

[1] Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A. *et al.* (2004) *Am J Hum Genet* **74,** 979-1000.

[2] Amundadottir, L.T., Sulem, P., Gudmundsson, J., Helgason, A., Baker, A. *et al.* (2006) *Nat Genet* 38, 652-658.

[3] G. Casella and R. Berger. Statistical Inference. Duxbury Press, 2001.

Supplementary Note 4

The -8 allele at DGS8737 can explain only a fraction of the admixture association

We set out to explore how much of the admixture signal could be explained by the -8 microsatellite allele identified by Amundadottir et al. [1].

For this analysis, we need estimates of the frequency of the -8 allele in European Americans and west Africans. We used f_{EW}=7.0% and f_{WA}=20.9%, based on genotyping in this study of 129 European American control chromosomes and 218 west African control chromosomes. The frequencies are fully consistent with the values reported in Amundadottir et al. [1].

We also need an estimate for the increased risk γ prostate cancer in African Americans per copy of the allele. This was estimated in Amundadottir et al. to be 1.60, but we have a much larger data set. Using the data reported in Table 3—from 966 African American cases age <72 and 797 African American controls—we estimate that the 95% credible interval for γ is 0.92-1.16. To be conservative in exploring how much of the admixture signal could be explained by the data, we focus on the highest value: γ=1.16.

How much of the admixture signal in younger African Americans can be explained by an allele with f_{EA} =7.0%, f_{WA} =20.9%, and γ=1.16? Defining A as the increased risk for prostate cancer due to an individual having an African-derived chromosome, we obtain:

$$A = \frac{\gamma f_{WA} + (1 - f_{WA})}{\gamma f_{EA} + (1 - f_{EA})} = 1.02.$$

This is far short of the 90% credible interval of 1.38-1.75 calculated in Supp. Note 2. The increased risk per copy would have to be γ=3.6 to fall within the credible interval. Thus, even if the microsatellite and admixture association are reflecting the causal variants (which is by no means guaranteed) the microsatellite is at best in weak linkage disequilibrium with the variants at 8q24 contributing to risk.

This analysis makes it clear that there are important risk allele(s) for prostate cancer that contribute to the admixture signal and that have not yet been identified.

References:

[1] Amundadottir, L.T., Sulem, P., Gudmundsson, J., Helgason, A., Baker, A. *et al.* (2006) *Nat Genet* 38, 652-658.