Genome-wide association study identifies novel loci predisposing to cutaneous melanoma

Christopher I. Amos1,*,†, Li-E Wang1,‡, Jeffrey E. Lee2,‡, Jeffrey E. Gershenwald2, Wei V. Chen1, Shenying Fang1, Roman Kosoy4, Mingfeng Zhang5, Abrar A. Qureshi5,6, Selina Vattathil1, Christopher W. Schacherer2, Julie M. Gardner2, Yuling Wang2, D. Tim Bishop7, Jennifer H. Barrett7, GenoMEL Investigators, Stuart MacGregor8, Nicholas K. Hayward8, Nicholas G. Martin8, David L. Duffy8, Q-Mega Investigators, Graham J. Mann9, Anne Cust10, John Hopper11, AMFS Investigators, Kevin M. Brown12, Elizabeth A. Grimm3, Yaji Xu1, Younghun Han1, Kaiyan Jing1, Caitlin McHugh13, Cathy C. Laurie13, Kim F. Doheny14, Elizabeth W. Pugh14, Michael F. Seldin4,†, Jiali Han5,6,15, and Qingyi Wei1,*,‡

1Department of Epidemiology, 2Department of Surgical Oncology and 3Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, 4Rowe Program in Genetics, University of California, UC Davis, Davis, CA, USA, 5Department of Dermatology and 6Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, 7Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, UK, 8Queensland Institute of Medical Research, Brisbane QLD 4029, Australia, 9Westmead Institute of Cancer Research, University of Sydney at Westmead Millennium Institute and Melanoma Institute Australia, Westmead, NSW 2145, Australia, 10Cancer Epidemiology and Services Research (CESR), Sydney School of Public Health, The University of Sydney, Sydney, Australia, 11Centre for Molecular, Environmental, Genetic and Analytic (MEGA) Epidemiology, Melbourne School of Population Health, University of Melbourne, Melbourne, Australia, 12Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, 13Department of Biostatistics, University of Washington, Seattle, WA, USA, 14Center for Inherited Disease Research, School of Medicine, Johns Hopkins University, Baltimore, MD, USA and 15Department of Epidemiology, Channing Laboratory, Harvard University, Boston, MA, USA

Received June 27, 2011; Revised and Accepted September 7, 2011

We performed a multistage genome-wide association study of melanoma. In a discovery cohort of 1804 melanoma cases and 1026 controls, we identified loci at chromosomes 15q13.1 (HERC2/OCA2 region) and 16q24.3 (MC1R) regions that reached genome-wide significance within this study and also found strong evidence for genetic effects on susceptibility to melanoma from markers on chromosome 9p21.3 in the p16/ARF region and on chromosome 1q21.3 (ARNT/LASS2/ANXA9 region). The most significant single-nucleotide polymorphisms (SNPs) in the 15q13.1 locus (rs1129038 and rs12913832) lie within a genomic region that has profound effects on eye and skin color; notably, 50% of variability in eye color is associated with variation in the SNP rs12913832. Because eye and skin colors vary across European populations, we further evaluated the associations of the significant SNPs after carefully adjusting for European substructure. We also evaluated the top 10 most significant SNPs by using data from three other genome-wide scans. Additional in silico data provided replication of the findings from the most significant region on chromosome 9p21.3.
1q21.3 rs7412746 \( (P = 6 \times 10^{-10}) \). Together, these data identified several candidate genes for additional studies to identify causal variants predisposing to increased risk for developing melanoma.

INTRODUCTION

Cutaneous melanoma (CM) is among the more common cancers, with more than 70,000 estimated new cases in the USA (excluding melanoma in situ) in 2011 (1). The risk for siblings of index cases who have CM is \( \sim 2\)-fold higher than that for the general population (2). Linkage studies and genome-wide association studies (GWASs) have identified several loci influencing CM risk. Cyclin-dependent kinase inhibitor 2A (CDKN2A or \( p16 \)), an alternate reading frame of CDKN2A (ARF or \( p14 \)), and cyclin-dependent kinase 4 (CDK4) in region 9p21.3 have been identified as high-penetrance CM susceptibility genes (3), but mutations in these genes are rare and explain only a small fraction of familial CM risk. However, GWASs found evidence that common variation in the CDKN2A region also commonly influence risk for melanoma (4,5). Mutations in the melanocortin 1 receptor (MC1R) gene in the chromosomal region 16q24.3 influence skin pigmentation, and a candidate gene analysis followed by several confirmatory studies documented an \( \sim 40\% \) increased risk of CM in those who carried the variants conferring lighter skin pigmentation according to specific mutations (6–8).

In addition, GWASs have identified several other loci that may contribute to CM risk. GWAS analyses have revealed several additional genes associated with skin pigmentation, including TYR (tyrosinase, chromosome 11q21) (9), TYRP1 (tyrosinase-related protein 1, chromosome 9p23) (10), SLC45A2 (solute carrier family 45, member 2, chromosome 5p13.3), SLC24A4 (solute carrier family 24, member 4, chromosome 14q32.12), and IRF4 (interferon regulatory factor 4, chromosome 6p25.3). A case–control study (11) that used a pooling approach to identify common variants influencing CM risk found several genetic factors influencing risk across a broad region of chromosome 20q11.2 that encompasses the genetic loci, PIGU, EIF2S2, NCOA6 and MYH17B. Finally, the locus PLA2G6 (4) on chromosome 22q13 was also previously associated with CM risk in a GWAS. The goal of the current GWAS was to identify novel variants associated with CM risk by using dense single-nucleotide polymorphism (SNP) arrays on a large population of CM cases that were collected uniformly from a single cancer center in the USA and to then validate these findings with use of additional data from US, Australian and European populations.

RESULTS

To identify risk variants for CM, we used an Illumina Omni1-Quad_v1-0_B array to genotype 1,016,423 SNPs in 3,115 participants. After the application of quality control (QC) criteria (Supplementary Material, Tables S1 and S2 and Fig. 1), genotypes were available for 1,804 Caucasian CM patients presenting to clinics (other than dermatology) at The University of Texas MD Anderson Cancer Center and 1,026 Caucasian cancer-free controls who were friends or acquaintances accompanying patients at their clinical visits. We analyzed 1,012,904 probes with a mean sample call rate of 99.66%. After applying QC filters, we retained data for analysis from 818,977 SNPs (818,237 autosomal or X chromosome SNPs and 740 pseudoautosomal SNPs) that had a minor allele frequency (MAF) of \( \geq 1\% \), were genotyped for 95% or more of retained participants and did not deviate from Hardy–Weinberg equilibrium \( (P > 1 \times 10^{-5}) \). Individual genotypes that had Illumina GenTrain version 1.0 quality scores of \( \leq 0.15 \) were set to missing. Figure 1 and Supplementary Tables 1 and 2 present an overall description of the sample processing and in silico replication phases of this study.

Initial analyses were performed using the additive logistic regression model implemented in PLINK (12). The Q–Q plots for the discovery sample are shown in Supplementary Material, Figures S1 and S2. The genomic inflation factor \( (\lambda) \) for the discovery sample was 1.02, and this was reduced when the first two principal components (PCs) were included as covariates \( (\lambda = 1.011) \). Association results across the genome after correcting for the two PCs are shown in Figure 2 and Supplementary Material, Table S3 (association tests with \( P < 1 \times 10^{-4} \)). The two regions reaching genome-wide significance \( (5 \times 10^{-8}) \) are located on 15q13.1 (HERC2/OCA2 region) and 16q24.3 (MC1R region) centered at or near the MC1R and HERC2/OCA2 genes, with a third highly significant region at chromosome 9p21.3 (near CDKN2A/ARF genes). In Table 1, we present the results for previously reported associations of CM risk with SNPs. Of the previously reported 27 risk-associated SNPs, 21 were significant in our study at \( P < 0.05 \). Individual SNPs that were previously associated with CM risk but did not reach significance in our study included rs1408799 in TYRP1, rs12896399 in SLC24A4, rs1800407 in OCA2, rs1805006 in MC1R, rs4911414 in LOC729547 and rs1015362 in EIF2S2, but associations for most previously reported SNPs were supported in our study, and except for the region around SLC24A4 on chromosome 14q32.12 there was at least one SNP in a region showing \( P < 1 \times 10^{-4} \).

To understand whether associations in regions that were previously reported to be associated with CM risk were likely to be due to effects from a single or multiple variants, we performed analyses in which the association of SNPs in a region was conditioned on the most significant SNP in that region. Conditioning on the most significant SNP in MC1R reduced the level of significance for all other SNPs, but some SNPs retained significance levels of \( 10^{-3} \) or higher (Supplementary Material, Fig. S3). Similarly, conditioning on the most significant SNP in the \( p16 \) region removed some, but not all, of the evidence for association in that region. These results suggest that multiple variants influence CM risk in these regions, or the SNP on which we conditioned may not be the causal variant, but was only in strong linkage.
disequilibrium (LD) with the causal variant, so that there remained some residual association. For 9p21.3 (CDKN2A region), associations were detected across a broad region that includes recombination hotspots even after conditioning on the most significant SNP, suggesting that multiple variants are involved.

**Figure 3** shows results from an association analysis of CM with SNPs in the HERC2/OCA2 region in samples from MD Anderson Cancer Center. The HERC2 SNP rs1129038 was the most significantly associated variant in the region \( P = 2.58 \times 10^{-8} \), odds ratio (OR) = 0.69, 95% confidence interval (CI) = 0.61–0.79). The next most significant genotyped SNP in this region was rs12913832 \( (P = 4.31 \times 10^{-8}; \text{OR} = 0.69; 95\% \text{CI} = 0.61–0.79) \). The variants rs1129038 and rs12913832 are in very high LD \( (r^2 = 98.5\%) \), but rs1129038 was not genotyped on other platforms used in subsequent studies; hence, further analyses were restricted to rs12913832.

The SNP rs1800407 in the flanking region of 15q13.1 (HERC2/OCA2 region) showed moderate evidence \( (P = 4.31 \times 10^{-8}) \) for association with CM risk in a candidate gene study but was not associated with CM risk in this study \( (P = 0.62, \text{OR} = 0.95; 95\% \text{CI} = 0.78–1.16) \). The most significant SNPs near the OCA2 gene yielded \( P \)-values \( (rs73377792, P = 6.3 \times 10^{-4}; rs4778138, P = 6.52 \times 10^{-4}) \) that were less extreme than those observed in the 15q13.1 (HERC2/OCA2) region in our study population. Further analyses conditioning on rs1129038 (Supplementary Material, Fig. S3) removed all remaining significant SNPs in the region.

Because the HERC2 gene was previously shown to be associated with pigmented phenotypes \( (13) \), we further evaluated the association of the rs12913832 SNP with pigmented phenotypes by using samples from MD Anderson (Supplementary Material, Table S4). Table 2 describes additional findings from the analysis of SNP rs12913832 in the MD Anderson participants. Results showed overwhelming evidence associating HERC2 rs12913832 with skin, eye and hair color phenotypes. Further analysis to evaluate the effect that this SNP had on these phenotypes showed that rs12913832 alone explained 50% of eye color (Supplementary Material, Table S5). In addition, we evaluated the association of HERC2 rs12913832 with pigmented phenotypes in 10,183 participants in the Nurses’ Health Study (NHS) and Health Professionals Follow-Up Study (HPFS). In these analyses, 567 cases and 7329 controls were genotyped using Illumina Bead arrays for HERC2 SNP rs12913832. We found extremely significant associations (Table 3). Stratifying by skin color, we observed strongest effects of the HERC2 SNP in lighter-skinned individuals \( (n = 1009, P = 0.0003, \text{OR} = 0.66, 95\% \text{CI} = 0.53–0.83) \), a
weaker effect in medium-skinned individuals \((n = 828, P = 0.02, OR = 0.77, 95\% CI = 0.61–0.96)\) and no effect in dark-skinned individuals \((n = 117, P = 0.75, OR = 1.10, 95\% CI = 0.61–2.00)\). In the Harvard cohort studies, the strongest effect was observed in light-skinned individuals \((n = 2173, P = 0.08 \text{ OR} = 0.80, 95\% CI = 0.63–1.03)\), and no effect was observed in medium-skinned \((n = 1916, P = 0.64, \text{ OR} = 1.05, 95\% CI = 0.85–1.32)\) or dark-skinned individuals \((n = 1916, P = 0.72, \text{ OR} = 0.94, 95\% CI = 0.67–1.32)\).

Because skin pigmentation and the HERC2 SNP frequency (Supplementary Material, Table S6) showed highly significant variability across European populations (14) and the CM risk is higher for northern European than for southern European populations, a potential concern can be raised that northern/southern European ancestry may be a confounder, causing a spurious association to be observed. We therefore derived the most likely European ancestry of the study participants.
Figure 4. PCA of CM cases and controls showing the relationship among populations of European and Middle-Eastern origin. (A) Analyses of 1804 CM cases and 1026 controls, with an additional 5001 Caucasians of known origin, are shown for PC1 and PC2 (A1, A2, A3), and for PC1 and PC3 (A4, A5, A6). The results are shown with either the known population groups color-coded (A1 and A4) or with controls color-coded (A2 and A5) or cases color-coded (A3 and A6). The population groups are indicated in the legend key. AJA 4GP are Ashkenazi Jewish Americans with four grandparents of Ashkenazi heritage. The group denoted ‘ALL’ are Americans of European ancestry. Results show similarity of PCA loadings for cases and controls in A2–A6. (B) The color code indicates the correspondence of controls or cases to seven different population groups with graphs for PC1 and PC2 (B1 and B3), and PC1 and PC3 (B2 and B4). The population groups were based on correspondence to PCA. The population groups corresponded to the range of the self-identified population group member defined by the first three PCs (see A). Where overlaps were present, the midpoints in the overlapped region were used to define the boundary between the population group assignment. The legend key shows the corresponding population groups with abbreviations for the following groups: Ashkenazi Jewish (ASHKEN), eastern European (EEUR), German (GER), Scandinavian (SCAN) and United Kingdom (UK). Analysis of CM case–control data alone yielded two PCs, whereas analysis including additional European controls yielded the additional third PC, shown in this figure.
CI = 0.97–1.04) and was not genotyped in the UK sample. The next most significant SNP in the US study, rs12913832, was not associated with CM risk in the Australian samples ($P = 0.63; OR = 0.98, 95\% CI = 0.89–1.07$) or UK samples ($P = 0.44; OR = 0.96; 95\% CI = 0.86–1.07$), but showed a trend towards association in the Harvard cohort studies ($P = 0.12; OR = 0.88, 95\% CI = 0.76–1.03$).

As shown in Figure 6, a region of chromosome 1q21.3 near the ARNT and LASS2 genes around 149.1 Mb was highly associated with CM risk in the samples from MD Anderson Cancer Center. The most significant SNP in the 1 Mb region including data from the other studies (Fig. 7) was rs7412746 ($P = 6.17 \times 10^{-10}$, OR = 0.88; 95\% CI = 0.84–0.91), which significantly replicated in all of the studies but Harvard: MD Anderson Cancer Center ($P = 0.003; OR = 0.85; 95\% CI = 0.76–0.94$), Australia ($P = 2.52 \times 10^{-7}; OR = 0.82; 95\% CI = 0.76–0.89$), GenoMEL ($P = 0.014; OR = 0.92; 95\% CI = 0.86–0.98$), but not in the smaller Harvard study ($P = 0.12; OR = 1.11; 95\% CI = 0.97–1.27$). The most significant SNP in the MD Anderson Cancer Center study, rs11204756 ($P = 3.61 \times 10^{-5}; OR = 1.26; 95\% CI = 1.13–1.40$) in the ANXA9 gene, had partially imputed data ($r^2 = 0.99$) by the Australian study ($P = 8 \times 10^{-4}; OR = 1.13; 95\% CI = 1.05–1.22$), yielding an overall association ($P = 3.59 \times 10^{-7}, OR = 1.17, 95\% CI = 1.10–1.24$) and is, therefore, worthy of future investigations.

DISCUSSION

This study provided confirmatory evidence of associations between genetic loci in several previously identified chromosome regions and CM risk, including the SLC45A2 region on chromosome 5p13.3, the 9p21.3 region encompassing CDKN2A, the region of TYR on chromosome 11q21, the region on chromosome 16q24.3 encompassing MC1R, a broad region on chromosome 20q11.2 and a region of 22q13 encompassing PLA2G6. Our study focused primarily on the discovery of novel variants influencing CM risk and we...
identified strong associations of SNPs in the HERC2/OCA2 region on chromosome 15q13.1 as well as a region of chromosome 1q21.3 that had not previously been well characterized. The HERC2 SNPs had previously been noted to associate strongly with pigmentation (5), but our finding is the first report that these SNPs influence CM risk in US populations. In the previous analysis of pigmentary phenotypes and SNPs in the HERC2/OCA2 region, the strongest association was found with SNPs rs12913832 and rs1129038 (13). These SNPs lie in a region that is in LD with both exons of HERC2 as well as the regulatory region of OCA2, so that further studies are needed to identify whether variation in HERC2 or OCA2 accounts for CM risk and pigmentary changes in this region. We also identified the SNP rs4778138 in OCA2, which showed a suggestive association with CM risk among the studies in which it was genotyped. These results implicate HERC2/OCA2 variation as a pigment-related gene that affects CM risk. Risk of CM could also be influenced by the effects of HERC2, which has recently been shown to play a role in ubiquitination after DNA damage (15) and to post-translationally regulate has recently been shown to play a role in ubiquitination and to post-translationally regulate.

Figure 7. Association tests from four studies using fixed-effects meta-analysis for a region of chromosome 1 spanning the ARNT/LASS2/ANXA9 region.

that has previously been associated with TH2-related responses that are induced during pemphigus (17). LASS2 is a ceramide synthase that has previously been shown to influence apoptosis after ionizing irradiation (18) and to play a role in tumor suppression in hepatocellular carcinoma cell lines (19). SETDB1 plays a role in trimethylation required for proviral silencing, which would seem a less likely candidate for influencing CM risk. ARNT is the aryl hydrocarbon nuclear translocator and plays a key role in regulating response to exogenous compounds that are catabolized by P450 enzymes. In summary, this study identified signals in several chromosomal regions and suggests that candidate genes including OCA2, HERC2, ARNT, ANXA9 and LASS2 warrant additional investigations.

MATERIALS AND METHODS

MD Anderson Cancer Center Study

The study participants for the discovery analysis were from a hospital-based case–control study of CM, for which cases were recruited from non-Hispanic white patients and controls at MD Anderson between March 1998 and August 2008. Samples and data were available from 931 CM patients and 1026 cancer-free controls (friends or acquaintances of patients reporting to other clinics), who were frequency-matched on age and sex, completed a comprehensive skin lifestyle questionnaire and passed QC filters for genotyping. This questionnaire was administered by an interviewer to 70% of patients and controls and was self-administered for the remaining 30%. An additional case series comprising 873 individuals presenting for the treatment for CM at MD Anderson was also included, bringing the total number of CM patients to 1804. The study protocols were approved by the Institutional Review Board at MD Anderson, and informed consent was obtained from all participants.

Genotyping and data QC

Tissue samples were collected as whole blood, using various DNA extraction methods (including Gentra, Qiagen and phenol/chloroform). DNA samples for the first-stage GWAS were genotyped with use of the Illumina HumanOmni1-Quad_v1-0_B array and were called using the BeadStudio algorithm, at the John Hopkins University Center for Inherited Disease Research (CIDR). We were able to satisfactorily genotype 1 012 904 of the 1 016 423 SNPs attempted (99.6%) with a mean sample call rate of 99.86% (Supplementary Material, Table S1). Supplementary Material, Table S1 summarizes a series of SNP filters applied to the original 1 140 419 SNPs and CNV probes. Before data release, 3519 SNPs failed the CIDR QC process with a missing call rate of >5%. SNPs with an MAF of ≤0.01, call rate of <95% or Hardy–Weinberg equilibrium in controls with a P-value of ≤10^-5 were excluded.

After the above criteria were applied, 818 237 genotyped autosomal or X chromosome SNPs and 740 pseudo-autosomal SNPs were available for the final association analysis. Imputation of ungenotyped SNPs in candidate regions (at least two genotyped SNPs in one region having P < 10^-4) was
performed for case and control samples with the use of the MACH program (20) using 1000 Genomes CEU (March 2010 release) as a reference panel. An additional 12 009 SNPs were imputed from those regions. In addition, genome-wide imputation using MACH and Hapmap2 Release 22 using the CEU population has been completed and data are available upon request for 2 373 692 SNPs with $r^2 > 0.8$, with an average quality score of 99.35%. For the most significant SNP in the HERC2 region, rs12913832, we performed confirmatory genotyping of 528 cases and controls using TaqMan. One subject was misclassified between the two platforms, one subject was genotyped as G/G by TaqMan and A/G by Illumina and one subject could not be genotyped by TaqMan but was found to have the G/G genotype by Illumina. Of the samples that were genotyped, 41 failed genotyping with a >10% missing rate across all SNPs; 11 samples had identity problems that could not be resolved. For this study, the identity-by-descent (IBD) coefficients were estimated using 116 002 autosomal SNPs in PLINK (12) and 5 unexpected duplicates and 15 related samples were removed. In total, 126 duplicated (67 expected duplicates), related (IBD) or outliers identified by PCA were excluded from the study. After these exclusions, 1952 cases and 1026 controls remained. Supplementary Material, Table S1 summarizes the whole-sample filters. From the total 2978 case and control subjects with data after QC, 138 in situ cases were removed from the study because they had indeterminate phenotypes; in addition, 10 patients with atypical melanocytic proliferation were excluded because they did not have invasive cancers. Data from 1804 cases and 1026 controls were analyzed for the association study of CM susceptibility.

**PCA and population structure of the MD Anderson sample**

We used an LD-pruned SNP set provided by the GENEVA (21) coordinating center for PCA adjustment to evaluate population structure. To select a set of SNPs for identifying population stratification, GENEVA thinned the initial SNPs to reduce LD to a set of 75 210 SNPs. SNPs were pruned using pairwise genotype correlation in PLINK. The pruning procedure consisted of two stages. In the first stage, short-range LD was removed. We used a window size of 50 SNPs with a 5 SNP offset and $r^2$ cut-off point of 0.2. For the second stage, we aimed to remove long-range LD. We used a window size of 150 SNPs with an offset of 5 Mb and $r^2$ cut-off point of 0.2. We adjusted for the first two components from PCA in our GWAS. No other PCs varied significantly between cases and controls. Those two eigenvectors were treated as covariables to adjust for population structure among study subjects. The GENEVA coordinating center provided an initial quality assessment of data and helped to organize the data for submission to dbGAP.

**Association analysis of the MD Anderson sample**

Association analysis with CM of genotyped SNPs or most likely genotypes from the imputation study was performed using the PLINK –logistic and –covar options (12). A logistic regression model was built to measure the additive effect of each SNP on susceptibility to CM. A likelihood ratio test was performed under the null hypothesis of $\chi^2$ distribution with one degree of freedom. The first two PCs were included to adjust for population structure. Q–Q plots portraying the associations of markers with CM risk are shown in Supplementary Material, Figure S1, including all of the directly genotyped markers, and in Supplementary Material, Figure S2 after excluding previously identified genes strongly associated with CM risk (MCIR and p16/arf). Significant regions with multiple SNPs that each had a $P$-value of $<10^{-4}$ were selected for further validation in other independent studies.

To evaluate associations with pigmentedary phenotypes, we regressed ordinal coding for hair color (1 = blonde, 2 = red, 3 = brown and 4 = black), eye color (1 = blue/grey, 2 = hazel/green and 3 = brown/black) and frequency of burns on a 1–4 scale (1 = always or usually burn, 2 = moderately burn, 3 = minimally burn and 4 = rarely or never burn). We used linear regression to test the association between minor allele counts and pigmentedary phenotypes, and the model was adjusted for age and sex.

**Genomic imputation**

MACH was used to impute ungenotyped SNPs in the candidate regions according to 1000 Genomes Project data. Based on the reference data with a denser set of markers, those untyped markers can be filled in by means of maximum-likelihood estimation. Imputation was run in one step since it was for small candidate regions. A total of 1682 SNPs with P-values of $<10^{-4}$ were found among 16 regions on 12 chromosomes, excluding chromosome 9 near the centromere where p16/arf/CDKN2A is located and chromosome 16 near MC1R, each of which contained at least two significant SNPs. Within a region of 50 kb from each side of these top SNPs, untyped genotypes were inferred with the use of MACH. Finally 12 009 SNPs were obtained with average $r^2 > 0.875$ for those candidate regions. The average posterior probability for the most likely genotype was 97.7%. As newer versions of the 1000 Genomes Project become available, better inference from it should be possible.

**Further adjustment for population structure in the MD Anderson sample**

Because skin pigmentation and CM risk show a North–South gradient across European populations, and allele frequencies in genes associated with skin pigmentation also show variation in allele frequencies, confounding between allele frequency and northern/southern European ancestry is a concern. We therefore performed additional studies to characterize the European ethnic background of study participants from MD Anderson and to allow for this inferred ancestry. Results, shown in Figure 4, demonstrate the inferred origins of study participants from MD Anderson according to the first two PCs. Supplementary Material, Table S7 displays results from analyses that were conditioned on the inferred most likely ancestral origin of each participant. The approach to selecting populations for characterizing European ancestry has been previously presented (22). Cases and controls that did not cluster within European ancestry groups were removed from analysis. In this analysis, 1795 cases and 931 controls were
retained, whereas an additional 1587 controls genotyped on 610K platforms that had genotypes close to the existing cases according to the first six PCs were used as additional controls in the analysis.

Materials and methods from Harvard cohort studies

We have conducted several GWASs on different disease outcomes [NHS: breast cancer by Illumina 550, and pancreatic cancer by Illumina 550; HPFS: advanced prostate cancer by Illumina 610; both NHS (part of the CGEMS) (23–25) and HPFS: coronary heart disease by Affymetrix 6.0, type 2 diabetes by Affymetrix 6.0, kidney stone by Illumina 610 and glaucoma by Illumina 660]. We included only controls in each study, except for the kidney stone study, in which we used both cases and controls. We excluded those with personal history of squamous cell carcinoma and basal cell carcinoma. Participants with CM diagnosis were the CM cases in this study, and those without CM diagnosis were the controls. In addition, we genotyped the rest of the CM cases identified in both cohorts who were not in these previous GWASs by Illumina 610. Finally, we included 494 CM cases and 5628 controls.

Based on the genotyped SNPs and haplotype information in the NCBI Build 35 of phase II Hapmap CEU data, we imputed genotypes for >2.5 million SNPs using the program MACH (20). Only SNPs with imputation quality $r^2 > 0.95$ and with MAF > 2.5% in each study were included in the final analysis. A total of 1 579 307 (NHS: 1 518 067; HPFS: 1 533 499) SNPs were included in the final meta-analysis of the NHS set and HPFS set.

Information on natural hair color at age 20 years and on childhood and adolescent tanning tendency was collected in both the NHS and HPFS prospective questionnaires, and information on natural eye color was collected in the HPFS only. The summary of the basic information of the five-component cohort studies used in the GWAS of pigmentation is shown in Supplementary Material, Table S4. The imputed dosage data on chromosome 15 were used for the association study. The imputed dosage for $r^2 > 0.95$ and with MAF > 2.5% in each study were included in the final analysis. A total of 1 579 307 (NHS: 1 518 067; HPFS: 1 533 499) SNPs were included in the final meta-analysis of the NHS set and HPFS set.

In silico replication studies

Confirmatory in silico replication was sought from GenoMEL, Harvard cohorts and the Australian consortium for the top 10 SNPs of significance for which another nearby SNP with LD $r^2 > 0.8$ also reached a significance level of $<10^{-4}$.

Australian sample

DNA was extracted from peripheral blood or saliva samples. Australian twin and endometriosis sample controls were genotyped at deCODE Genetics (Reykjavik, Iceland) on the Illumina HumanHap610W Quad and Illumina HumanHap660 Quad Beadarrays, respectively. AMFS controls were genotyped by Illumina (San Diego) on Illumina Omni1-Quad arrays. Cases were genotyped by Illumina (San Diego) on Illumina Omni1-Quad (568 AMFS cases, 699 Q-MEGA cases) and HumanHap610W Quad arrays (998 Q-MEGA cases). All genotypes were called with the Illumina BeadStudio software. SNPs with a mean BeadStudio GenCall score $<0.7$ were excluded from the control data sets. All samples had successful genotypes for $>95\%$ of SNPs. SNPs with call rates either $<0.95$ (MAF > 0.05) or $<0.99$ (MAF > 0.01), Hardy–Weinberg equilibrium in controls $P < 10^{-6}$ and/or MAF < 0.01 were excluded. Cryptic relatedness between individuals was assessed through the production of a full identity-by-state matrix. Ancestry outliers were identified by PC analysis, using data from 11 populations of the HapMap 3 Project and 5 northern European populations genotyped by the GenomeUttwin consortium, using the EIGENSOFT package. Individuals lying $\geq2$ standard deviations from the mean PC1 and PC2 scores were excluded from subsequent
analyses. Following these exclusions, there were 2168 case samples (1242 typed on the Omni1-Quad and 926 typed on the Hap610 arrays, respectively) and 4387 controls (431 typed on the Omni1-Quad and 3956 typed on the Hap610 arrays) retained for subsequent analyses (Table 1). Individuals typed on the Omni1-Quad array had genotypes for up to 81616 SNPs, whereas individuals typed on the Hap610/670 arrays had genotypes for up to 544483 SNPs. There were 299394 SNPs passing QC and overlapping between these arrays (and hence directly genotyped on all Australian samples).

Imputation for the Australian samples was performed using MACH (20) with 1000 Genomes Project [June 2010 release] data obtained from people of northern and western European ancestry collected by the Centre d’Etude du Polymorphisme Humain. Imputation was based on a set of autosomal SNPs common to all CM case–control samples (n = 292043). Imputation was run in two stages. First, data from a set of representative Australian sample individuals were compared with the phased haplotype data from the 1000 Genomes data to generate recombination and error maps. For the second stage, data were imputed for all individuals using the phased 1000 Genomes data as the reference panel and the recombination as well as error files generated in the first stage. In total, 5480804 1000 Genomes SNPs could be imputed with imputation r^2 > 0.5.

Association analysis of genotyped SNPs was performed using the PLINK –assoc option (12). Analysis of dosage scores from the imputation analysis was done using mach2dat.

GenoMEL sample

The GenoMEL data incorporated in the in silico replication came from a GWAS of 2804 cases and 1835 controls for CM collected at 11 different centers across Europe and Israel. These were genotyped in two phases, phase 1 on the Illumina HumanHap300 BeadChip version 2 duo array (317K SNPs) and phase 2 on the Illumina HumanHap1.2 million array (610K SNPs). These are supplemented by 3878 controls from the WTCCC study [WTCCC (26)] genotyped on the Illumina HumanHap1.2 million array and 1905 French controls genotyped by Centre National de Genotypage on the Illumina HumanHap1.2 million array. Individuals were excluded for low call rate (<97% on the array on which the sample was genotyped), non-European ethnicity (determined by PCA), sex discrepancy with recorded phenotype information, first-degree or closer relatedness with another sample or recommendation for exclusion by WTCCC (26). QC was applied to SNPs separately for each genotyping platform. SNPs were excluded for a low call rate (<97%), Hardy–Weinberg equilibrium P-value of <10^−20 or recommendation for exclusion by WTCCC (usually on the basis of poor clustering). A trend test was applied to each SNP in turn stratified by broad geographic region (eight regions pre-specified).

WEB RESOURCES

EIGENSOFT: http://genepath.med.harvard.edu/~reich/Software.htm.
REFERENCES


APPENDIX

MEMBERS OF THE GENOMEL COLLABORATION


Brisbane: The Queensland study of Melanoma: Environmental and Genetic Associations (Q-MEGA) Principal Investigators: Nicholas G. Martin, Grant W. Montgomery, David L. Duffy, David C. White, Dennis Town, Stuart MacGregor and Nicholas K. Hayward.

The Australian Cancer Study (ACS) Principal Investigators: David Whitman, Penny Webb, Adele Green, Peter Parsons, David Purdie and Nicholas Hayward.

Emilia-Romagna: Maria Teresa Landi, Donato Calista, Giorgio Landi, Paola Minghetti, Fabio Arcangelo and Pier Alberto Bertazzi.


Glasgow: Rona Mackie, Julie Lang.


Norway: Oslo University Hospital: Per Helsing, Per Arne Andersen, Helge Rootwelt.


Poland: Tadeusz Dębniak, Jan Lubiiński, Aneta Mirecka and Sławomir Ertmański.

Slovenia: Szrgi Novakovic, Marko Hocevar, Barbara Peric and Petra Cerkovenk.

Stockholm: Veronica Höiöm and Johan Hansson.


See also: www.GenomEL.org.