Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma

Harvind S. Chahal1,*, Wenting Wu2,*, Katherine J. Ransohoff1, Lingyao Yang3, Haley Hedlin3, Manisha Desai3, Yuan Lin2, Hong-Ji Dai2,4, Abrar A. Qureshi5,6,7, Wen-Qing Li5,6, Peter Kraft8,9, David A. Hinds10, Jean Y. Tang1,**, Jiali Han2,4,8,** & Kavita Y. Sarin1,**

Basal cell carcinoma (BCC) is the most common cancer worldwide with an annual incidence of 2.8 million cases in the United States alone. Previous studies have demonstrated an association between 21 distinct genetic loci and BCC risk. Here, we report the results of a two-stage genome-wide association study of BCC, totalling 17,187 cases and 287,054 controls. We confirm 17 previously reported loci and identify 14 new susceptibility loci reaching genome-wide significance (P < 5 × 10⁻⁸, logistic regression). These newly associated SNPs lie within predicted keratinocyte regulatory elements and in expression quantitative trait loci; furthermore, we identify candidate genes and non-coding RNAs involved in telomere maintenance, immune regulation and tumour progression, providing deeper insight into the pathogenesis of BCC.
With 2.8 million new cases diagnosed annually in the United States, basal cell carcinoma (BCC) is the most common cancer worldwide and contributes substantially to morbidity. BCC risk has been associated with ultraviolet (UV) exposure, fair skin, arsenic exposure, ionizing radiation, chronic immunosuppression, male gender and age. Since 2008, six population-based genome-wide association studies (GWAS) of BCC have been reported identifying 16 regions of susceptibility. Candidate gene studies have identified five additional pigmentation loci associated with BCC, including IRF4, SLC45A2, RALY, TYR and OCA2. Here, we report the largest-to-date two-stage genome-wide association meta-analysis for BCC totalling 17,187 cases and 287,054 controls. The results of this study confirm 12 of 16 loci from prior GWAS, along with 5 of 5 loci from previous candidate gene studies, and identify 14 novel susceptibility loci for BCC.

**Results**

Stage 1 consisted of 12,945 self-reported BCC cases and 274,252 controls of European ancestry from 23andMe research participants (Supplementary Table 1). Validation of self-reported surveys with adjudicated medical records revealed a sensitivity and specificity of 93% and 99%, respectively (Supplementary Table 2), indicating a misclassification rate of less than 10%. Simulation analysis demonstrated that this misclassification rate only modestly reduced the power to detect associations in this study (Supplementary Fig. 1). Stage 2 consisted of an independent GWAS cohort of 4,242 BCC cases and 12,802 controls of European ancestry from the Nurses’ Health Study and Health Professionals Follow-Up Study (Supplementary Table 1). Subsequently, meta-analysis of stages 1 and 2 was performed, encompassing 17,187 cases and 287,054 controls (Supplementary Table 1). Further information on methodology and imputation quality control is presented in the Online Methods, Supplementary Tables 3–5, and Supplementary Figs 2–7.

**Stage 1 analysis.** Twenty-eight index single nucleotide polymorphisms (SNPs) were associated with BCC at the genome-wide significance level ($P < 5.0 \times 10^{-8}$, logistic regression) in stage 1 and are depicted in the Manhattan plot (Fig. 1). Subset analysis revealed relatively consistent effect sizes across age and gender for these 28 SNPs (Supplementary Tables 6–7, Supplementary Fig. 8).

Interestingly, slightly larger effect sizes tended to occur in younger cases, suggesting that other risk factors may play an increasing role with age. As 10% of the BCC cases in our stage 1 cohort subsequently developed melanoma, we also investigated whether the co-occurrence of melanoma contributed to the observed associations with BCC risk. We therefore computed association tests for the stage 1 index SNPs in BCC cases with and without melanoma. All 28 SNPs displayed consistent effect sizes across the two groups (Table 1, Supplementary Fig. 9), indicating that they are independently associated with BCC susceptibility.

**Stage 2 and combined meta-analysis.** Twenty of the 28 index SNPs were replicated in the stage 2 analysis ($P < 0.05$, logistic regression). While some loci did not reach statistical significance in stage 2, their 95% confidence intervals (for odds ratios) overlapped with the corresponding stage 1 confidence intervals. Meta-analysis of stages 1 and 2 identified a total of 31 loci reaching genome-wide significance ($P < 5 \times 10^{-8}$, logistic regression) for association with BCC. Among these 31 loci, 17 are previously reported (Supplementary Tables 8–12). The remaining 14 are novel BCC susceptibility loci (Table 2, Supplementary Table 13). Forest plots and regional association plots for these 14 SNPs are provided in Supplementary Figs 10–12. Of these 14 novel risk variants, 10 reached genome-wide significance in stage 1 and 4 reached genome-wide significance in the combined meta-analysis. As many pigmentation loci have been associated with BCC susceptibility, we adjusted for pigmentation phenotype in our stage 2 cohort and did not observe a significant difference between adjusted and unadjusted results for the 14 novel risk variants (Supplementary Table 14).

**Heritability of BCC and gene expression analysis.** To measure the proportion of BCC heritability that can be attributed to these SNPs, we calculated the familial relative risk for BCC as outlined by the Cancer Oncological Gene-Environment Study. Overall, 10.98% of familial relative risk for BCC is explained by the 31 genome-wide significant loci; of this percentage, the 14 novel susceptibility loci account for 2.62%. To further explore the role of these 14 loci in BCC pathogenesis, we evaluated expression levels of nearby genes in BCC tissue using two data sets from the Gene Expression Omnibus (GEO) (GSE53462 and GSE7553). Three loci demonstrated significant differential gene expression.
Ultraviolet exposure analysis). CASC15 6p22.3 rs2294214 6p22.3 HLA-DQA2 rs7776701 7p12.3 PLIN3 ZBTB10 rs11993814 8q21.13 LPP rs191177147 3q28 rs10810657 9p22.2 BNC2 rs10093547 11p15.5 ZFHX4 rs191177147 LPP -0.114 0.016 6.3 × 10^-13 -0.009 0.057 8.8 × 10^-11 0.577 rs9275642 HLA-DQB1/DQA2 rs0.123 0.019 2.4 × 10^-11 -0.080 0.065 2.2 × 10^-11 0.346 rs7874604 CSDKN2 rs0.098 0.015 1.6 × 10^-10 0.095 0.055 8.5 × 10^-12 0.591 rs11170164 KRT5 rs0.163 0.025 1.2 × 10^-10 0.134 0.094 1.6 × 10^-11 0.303 rs73186343 CUX1 rs0.109 0.017 1.1 × 10^-10 0.067 0.060 2.6 × 10^-11 0.391 rs11993814 ZBTB10 rs0.097 0.016 2.1 × 10^-9 -0.146 0.058 1.2 × 10^-2 0.413 rs2776353 LINCO0111 rs0.092 0.015 1.5 × 10^-9 -0.063 0.055 2.5 × 10^-11 0.623 rs1765871 ATPIA1 rs0.068 0.014 1.1 × 10^-6 -0.145 0.052 4.8 × 10^-3 0.981 rs10425559 TICAM1/PLIN3 rs0.086 0.014 2.7 × 10^-9 0.004 0.052 9.3 × 10^-10 0.458 rs1050529 HLA-B rs0.106 0.019 2.5 × 10^-8 -0.096 0.069 1.7 × 10^-10 0.486 rs7776701 TNS3 rs0.081 0.014 6.9 × 10^-9 0.041 0.050 4.1 × 10^-11 0.711 rs78097823 TTC28 rs0.164 0.034 2.4 × 10^-6 0.354 0.120 3.5 × 10^{-3} 0.466

Results generated from logistic regression models fit separately in melanoma controls (N) and cases (Y). Includes P-value (Int P) from a likelihood ratio test for adding an interaction with melanoma status to the GWAS model. Of the 274,252 BCC cases, 3138 were melanoma cases and 268,282 were melanoma controls. Of the 12,945 BCC cases, 1,350 were melanoma cases and 11,465 were melanoma controls.

| SNP | Region | Gene | Maj/min | MAF (avg imp r²) | P | OR | P | OR |
|-----|--------|------|---------|-----------------|---|----|---|----|
| rs2116709 | 3p13 | FOXPI | A/T | 0.40 (0.91) | 7.9 × 10^-15 | 0.89 | 6.1 × 10^-4 | 0.91 | 2.3 × 10^-17 |
| rs10810657 | 9p22.2 | BNC2 | A/T | 0.41 (0.98) | 5.1 × 10^-14 | 0.90 | 5.7 × 10^-5 | 0.90 | 1.5 × 10^-12 |
| rs191177147 | 3q28 | LPP | G/T | 0.39 (0.80) | 3.2 × 10^-12 | 1.11 | 1.0 × 10^-3 | 1.10 | 1.2 × 10^-9 |
| rs9275642 | 6p21.32 | HLA-DQA2 | C/T | 0.21 (0.89) | 1.2 × 10^-11 | 0.89 | 2.7 × 10^-2 | 0.81 | 2.4 × 10^-12 |
| rs73186343 | 7q11.23 | CUX1 | G/A | 0.24 (0.96) | 8.5 × 10^-11 | 0.90 | 2.3 × 10^-4 | 0.89 | 1.5 × 10^-13 |
| rs11993814 | 8q21.13 | ZBTB10 | C/T | 0.26 (1.0) | 2.8 × 10^-10 | 0.91 | 4.5 × 10^-2 | 0.94 | 8.8 × 10^-11 |
| rs2776353 | 2q22.3 | LINCO0111 | A/T | 0.33 (0.96) | 5.0 × 10^-10 | 0.91 | 7.7 × 10^-4 | 0.91 | 1.6 × 10^-12 |
| rs10425559 | 19p13.3 | PLIN3 | G/A | 0.40 (0.97) | 3.8 × 10^-11 | 0.92 | 8.4 × 10^-1 | 0.99 | 2.8 × 10^-9 |
| rs1050529 | 6p21.33 | HLA-B | C/T | 0.25 (0.71) | 4.6 × 10^-9 | 0.90 | 2.7 × 10^-11 | 0.89 | 2.6 × 10^-9 |
| rs7776701 | 7p12.3 | TNS3 | C/T | 0.48 (0.98) | 5.3 × 10^-8 | 0.93 | 5.1 × 10^-1 | 0.94 | 4.2 × 10^-8 |
| rs9267650 | 6p22.3 | NEU | A/T | 0.05 (0.98) | 2.4 × 10^-8 | 1.18 | 2.0 × 10^-1 | 1.09 | 1.1 × 10^-8 |
| rs7907606 | 10q24.3 | OFBC1 | T/G | 0.17 (0.96) | 7.4 × 10^-8 | 1.10 | 2.4 × 10^-2 | 1.08 | 4.7 × 10^-9 |
| rs4710154 | 6q27 | MIR3939 | A/T | 0.32 (0.93) | 8.1 × 10^-8 | 1.03 | 4.3 × 10^-2 | 1.06 | 1.1 × 10^-8 |
| rs2294214 | 6p22.3 | CASC15 | A/C | 0.32 (0.95) | 2.6 × 10^-5 | 1.06 | 5.9 × 10^-5 | 1.13 | 3.1 × 10^-8 |

SNPs that met genome-wide significance (P < 5 × 10^-8, via logistic regression) in overall meta-analysis are listed; these SNPs have not been associated with BCC in previous GWAS reports. Additionally, we report genetic loci, nearest genes, major allele, minor allele, minor allele frequency (MAF) as calculated from stage 1 data, average imputation r² (a measure of imputation quality) for stage 1, and odds ratio (OR) with P-value for each stage, calculated with respect to the minor allele. In stage 1, we analysed 12,945 BCC cases and 274,252 controls. Stage 2 included 4,242 BCC cases and 12,802 controls from NHS. We then combined data from stage 1 and stage 2 (which resulted in 17,187 BCC cases and 287,054 controls) and performed fixed-effect meta-analysis. Statistics for effect heterogeneity (Q and I²) are included in Supplementary Table 13. All subjects were from the US and of European ancestry.

Genotyping results in stage 2.

The results generated from logistic regression models fit separately in melanoma controls (N) and cases (Y). Includes P-value (Int P) from a likelihood ratio test for adding an interaction with melanoma status to the GWAS model. Of the 274,252 BCC cases, 3138 were melanoma cases and 268,282 were melanoma controls. Of the 12,945 BCC cases, 1,350 were melanoma cases and 11,465 were melanoma controls.

in BCC relative to normal skin: 3p13 FOXPI, 7p12.3 TNS3 and 6p22.3 CASC15 (Fig. 2, P < 0.05, linear models for microarray analysis).

Gene-environment interaction analysis. Ultraviolet exposure and pigmentation phenotypes have been associated with BCC risk and may interact with genetic variants to confer BCC susceptibility. To further explore the influence of such factors on our study, we tested for interactions between all 31 significant loci and UV exposure, hair colour, number of sunburns and tanning ability (Supplementary Table 15). This analysis revealed only one significant interaction, between rs191177147 (LPP) and hair colour. Additional testing in three separate hair colour
groups demonstrated that rs191177147 significantly interacted with the light brown and dark brown/black hair groups (Table 3, \( P = 2.9 \times 10^{-6} \), logistic regression).

**Discussion**

The 14 novel susceptibility loci cluster into five functional categories: telomere biology, immune regulation, tumour progression, non-coding RNA and pigmentation.

Variants within the TERT locus have previously been associated with BCC susceptibility, thus implicating telomere maintenance in BCC development. Here, we identify a novel BCC susceptibility locus, OBFC1, that is also involved in telomere maintenance. rs7907606 at 10q24.3 (\( P = 4.7 \times 10^{-5} \), odds ratio \( OR = 1.10 \), logistic regression) lies 3 kb upstream of the OBFC1 gene, a member of the heterotrimer CST complex. This complex restricts telomere extension by binding telomeric DNA and the ability of other proteins to recruit telomeres. Variants in the OBFC1 locus have been associated with mean leukocyte telomere length. rs7907606 is in linkage disequilibrium (LD) with four of these variants (Supplementary Table 16), including rs4387287 (\( r^2 = 0.98 \), \( D^2 = 1.00 \), which lies within the promoter of OBFC1 and is an eQTL for this gene in sun-exposed skin (\( P = 8.3 \times 10^{-6} \), GTEx V6 analysis). Though widely studied, the connection between telomere length and cancer is context dependent, as both long and short telomeres are linked to malignancy. This discrepancy is readily apparent in the context of skin cancer, where longer telomeres are associated with melanoma and shorter telomeres are associated with BCC. Our results further implicate telomere homeostasis in BCC pathogenesis.

We also identified six novel susceptibility loci associated with immune regulation. Two SNPs, rs1050529 and rs9275642, are in human leukocyte antigen (HLA) regions. rs1050529 at 6p21.33 (\( P = 2.6 \times 10^{-9} \), \( OR = 0.90 \), logistic regression) lies within an exon of HLA-B, and leads to a non-conservative amino acid substitution (A65T). rs1050529 has predicted enhancer and promoter activity in keratinocytes and is an eQTL in six tissues. Variants in HLA-B have been associated with a range of autoimmune conditions, including vitiligo and psoriasis. HLA-B encodes the heavy chain component of major histocompatibility complex (MHC) class I molecules, which present endogenously synthesized peptides to cytotoxic T cells. Downregulation of MHC class I molecules is a feature of many types of cancers, which is thought to enable tumour cells to evade recognition and destruction by T cells. An analysis of 91 human melanoma cell lines revealed decreased expression of class I molecules in 67% of the cell lines, with HLA-B being the most common. The second HLA SNP reaching genome-wide significance was rs9275642 at 6p21.32 (\( P = 2.4 \times 10^{-12} \), \( OR = 0.89 \), logistic regression), which is located 24 kb upstream of HLA-DQA2. This SNP is in tight LD with rs9275640 (\( r^2 = 0.89 \), \( D^2 = 0.99 \), which is an eQTL for this gene in sun-exposed skin (\( P = 3.6 \times 10^{-6} \), GTEx V6 analysis). HLA-DQA2 variants are also associated with autoimmune diseases, including type 1 diabetes, rheumatoid arthritis and alopecia areata. These findings represent the first genome-wide significant association between MHC genes and BCC risk.

The third immune-related SNP, rs191177147 at 3q28 (\( P = 9.8 \times 10^{-17} \), \( OR = 1.11 \), logistic regression), resides within an intron of LPP and is in LD with rs1464510 (\( r^2 = 0.54 \) and rs9860547 (\( r^2 = 0.68 \), Supplementary Table 17); the former is associated with autoimmune diseases such as celiac disease, rheumatoid arthritis, juvenile idiopathic arthritis and vitiligo, while the latter is associated with allergy. LPP encodes an intracellular protein that shuttles between the cell membrane and the nucleus, where it interacts with transcription factors to modulate gene expression. LPP overexpression has been reported in squamous cell lung carcinomas and primary sarcomas. We also found an association between rs191177147 and hair colour, suggesting that this SNP may also contribute to BCC risk by altering pigmentation phenotypes.

Another SNP with potential significance in immune regulation is rs10425559 at 19p13.3 (\( P = 2.9 \times 10^{-9} \), \( OR = 0.93 \), logistic regression), which is intergenic and flanked by TICAM1 and PLIN3. It is linked to rs7255265 (\( r^2 = 0.68 \), \( D^2 = 0.9 \), Supplementary Table 18), which is located in an exon of TICAM1 and has predicted enhancer activity in keratinocytes. TICAM1 is an intracellular toll-like receptor (TLR) adaptor molecule involved in innate immunity; moreover, it acts as a pro-apoptotic tumour suppressor by mediating the interaction between TLR-3 and caspase-8 in some malignancies, including melanoma. PLIN3 encodes a cytosolic protein that binds to the GTPase RAB9, a member of the RAS oncogene family. Overexpression of PLIN3 has been linked to cervical carcinoma.
The fifth immune-related SNP, rs9267650 at 6p21.3 (P = 1.1 × 10⁻⁸, OR = 1.17, logistic regression), lies 0.5 kb downstream of NEUI, which encodes a lysosomal enzyme implicated in many diverse processes, including activation of TLRs, wound healing, and suppression of ovarian carcinoma. Finally, rs11993814 at 8q21.13 (P = 1.1 × 10⁻¹², OR = 0.91, logistic regression), located 9 kb upstream of ZBTB10, is an eQTL in two tissues and is in LD with rs6998967 (r² = 0.6, D′ = 1.0), associated with late-onset myasthenia gravis. ZBTB10 encodes a zinc finger transcription factor involved in regulating the expression of Interleukin-10 through suppression of Sp1. Abnormal interaction between ZBTB10 and Sp1 is seen in several different cancer cell lines, with ZBTB10 consistently exhibiting tumour-suppressing activity. All together, these findings implicate a number of immune regulatory loci in BCC susceptibility.

Four of our novel susceptibility loci are associated with tumour progression. rs2116709 at 3p13 (P = 5.7 × 10⁻¹⁶, OR = 0.91, logistic regression) resides within an intron of FOXP1, a transcription factor that, in addition to regulating organ development, acts as a tumour suppressor in some cancers (for example, prostate) and as an oncprotein in others (for example, oesophagus). It is overexpressed in oesophageal cancer and many types of lymphomas, including cutaneous B-cell lymphomas. Sp11–31. Abnormal interaction between ZBTB10 and Sp1 is significantly upregulated in BCC.

Interestingly, two of the 14 novel susceptibility variants reside near or within long non-coding RNA genes. rs2776353 at 21q22.3 (P = 2.0 × 10⁻¹⁴, OR = 0.91, logistic regression) is located 10 kb upstream of LINC00111, while rs2294214 (P = 3.1 × 10⁻⁸, OR = 1.07, logistic regression) lies within an intron of CASC15 (also known as LINC00340); both SNPs have predicted enhancer and promoter activity in keratinocytes. A recent study compared the expression of long non-coding RNAs in BCC samples to that of normal skin controls and found that CASC15 was overexpressed in BCC. Overexpression of CASC15 has also been implicated in melanoma progression and metastasis. Our analysis of independent BCC expression data confirmed significant upregulation of CASC15 in BCC, further implicating this gene in BCC development.

In addition to confirming the previously reported association of five pigmentation loci with BCC, we identified a novel susceptibility locus—rs10810657 at 9p22.2 (P = 1.5 × 10⁻¹⁷, OR = 0.90, logistic regression)—that may also influence pigmentation. rs10810657, located 14 kb upstream of BNC2, reached genome-wide significance in the meta-analysis and is an eQTL for BNC2 in blood (Table 2). This SNP is in LD with rs12350739 (r² = 0.87, D′ = 0.99, Supplementary Table 20), associated with human skin pigmentation via regulation of BNC2 transcription in melanocytes, and rs62543565 (r² = 0.7, D′ = 0.87), associated with the age-related development of facial pigmented spots. rs10810657 is also linked to rs2153271 (r² = 0.9, D′ = 0.99), which is intrinsic to BNC2 and associated with freckling. BNC2 codes for basonuculin 2, a protein thought to function both as an RNA-processing enzyme and a transcription factor. It is expressed in melanocytes and, to a lesser extent, keratinocytes, with higher expression levels corresponding to darker skin pigmentation.

In summary, this two-stage meta-analysis represents the largest GWAS for BCC and identified 14 novel susceptibility loci with roles in telomere maintenance, immune regulation and tumour progression. Further investigation of these loci will improve our understanding of BCC pathogenesis and improve our ability to prevent these common tumours.

Methods

Stage 1 study design and population. 23andMe Inc. (Mountain View, CA), a genetics company, provided free access to anonymized genetic and phenotypic information for stage 1 of this GWAS. All information came from 23andMe research participants who provided informed consent to participate in research, in accord with 23andMe’s human subjects protocol (reviewed and approved by Ethical and Independent Review Services, an AAHRPP accredited IRB). 23andMe gathers genetic information by genotyping sample material provided by its research participants; phenotypic information is collected via research participant responses to online surveys. Inclusion and exclusion criteria are discussed below. Sample sizes for cases and controls were not predetermined as the intention was to maximize the number of subjects in both groups; hence, all subjects passing inclusion and exclusion criteria were included, which resulted in 12,945 BCC cases and 274,252 controls.

Stage 1 genome-wide association analysis. Association analysis for stage 1 was performed using logistic regression, assuming an additive model for allelic effects. The analysis was adjusted for age, sex and population stratification (using the first five principal components), generating the following model:

\[
\text{BCC diagnosis} \sim \text{age} + \text{sex} + \text{pop} + \text{pc1} + \text{pc2} + \text{pc3} + \text{pc4} + \text{genotype}
\]

The association test P-value was computed using a likelihood ratio test. Results for the X chromosome were computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele. Additionally, test statistics were adjusted for genomic control to correct for residual population stratification persisting after principal component analysis; the genomic control inflation factor was 1.085 (computed from the median P-value for results that passed quality control).

Genome-wide association analysis generated a set of index SNPs. The index SNPs show information for the most-associated SNP in each associated region. We
define regions of interest by identifying SNPs with $P < 10^{-5}$, then grouping these into intervals separated by gaps of at least 250 kb, and choosing the SNP with smallest $P$ within each interval.

Stage 1 genotyping and quality control. Samples were genotyped on one of four genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550 þ BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The V3 platform was the Illumina OmniExpress þ BeadChip, with custom content to improve the overlap with our V2 array, with a total of about 950,000 SNPs. The V4 platform in current use is a fully custom array, including a lower redundancy subset of V2 and V3 SNPs with additional coverage of lower-frequency coding variation, and about 570,000 SNPs. Samples that failed to reach 98.5% call rate were re-analysed. Individuals whose analyses failed repeatedly were re-contacted by 23andMe to provide additional samples, as is done for all 23andMe research participants.

Individuals were only included if they had >97% European ancestry, as determined through an analysis of local ancestry\(^5\). Briefly, this analysis first partitions phased genomic data into short windows of about 100 SNPs. Within each window, a support vector machine is used to classify individual haplotypes into one of 31 reference populations. The support vector machine classifications are then fed into a hidden Markov model (HMM) that accounts for switch errors and incorrect assignments, and gives probabilities for each reference population in each window. Finally, simulated admixed individuals are used to recalibrate the HMM probabilities so that the reported assignments are consistent with the simulated admixture proportions. The reference population data are derived from public data sets (the Human Genome Diversity Project, HapMap and 1000 Genomes), as well as 23andMe research participants who have reported having four grandparents from the same country.

A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm\(^6\). Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments identical-by-descent. This level of relatedness (roughly 20% of the genome) corresponds approximately to the minimal expected sharing between first cousins in an outbred population.

Participant genotype data were imputed against the March 2012 v3 release of 1000 Genomes reference haplotypes\(^7\). Data for each genotyping platform were phased and imputed separately. First, Beagle52 (version 3.3.1) was used to phase batches of 8000–9000 individuals across chromosomal segments of no more than 2000 SNPs. SNPs with Hardy-Weinberg equilibrium $P < 10^{-20}$, call rate <95%, or with large allele frequency discrepancies compared to European 1000 Genomes reference data were excluded. Frequency discrepancies were identified by computing a $2 \times 2$ table of allele counts for European 1000 Genomes samples and 2000 randomly sampled 23andMe research participants with European ancestry, and identifying SNPs with a chi-squared $P < 10^{-15}$. Each phased segment was imputed against all-ethnicity 1000 Genomes haplotypes (excluding monomorphic and singleton sites) using Minimac2\(^8\), using 5 rounds and 200 states for parameter estimation.

For the non-pseudoautosomal region of the X chromosome, males and females were phased together in segments, treating the males as already phased; the pseudoautosomal regions were phased separately. Males and females were then imputed together using minimac, as with the autosomes, treating males as homzygous pseudo-diploids for the non-pseudoautosomal region.

Stage 1 subset analyses. Subset analysis by age and gender was performed for the genome-wide significant index SNPs in stage 1. For age-based analysis, the stage 1 cohort was divided into four age intervals with similar effective sample sizes based on principal component analysis control sample counts. Stage 1 association test results were then computed within each of these age intervals for the 28 SNPs. The interaction between genotype effect and age interval was also calculated. For all these association tests, we used the same covariates used in stage 1: age, sex and five principal components. Thus, association tests within a specific age interval were still adjusted for age as a continuous covariate. For gender analysis, we compared effect sizes estimated in men versus effect sizes estimated in women for the 28 SNPs. We also performed logistic regression separately in the male and female subsets, and calculated $P$-values from a likelihood ratio test for adding a gender by genotype interaction to the full logistic regression models. For melanoma subset analysis, association tests for BCC were computed separately in melanoma controls and melanoma cases.

Stage 1 phenotype categorization. 23andMe identified BCC cases by using research participants’ self-reported answers to online questionnaires. Subjects who answered ‘Yes’ and/or selected BCC from a dropdown menu in response to at least one of the following questions were defined as cases: ‘Have you ever been diagnosed by a doctor with basal cell carcinoma?’, ‘What type of skin cancer did you have? Please check all that apply.’, ‘What type of skin cancer or cancers have you been diagnosed with? Please check all that apply.’, ‘Have you ever been diagnosed with basal cell carcinoma?’ ‘Have you ever been diagnosed or treated for any of the following conditions?’ Controls were defined as subjects who answered ‘No’ and did not select BCC from any relevant dropdown menus. In addition, subjects who answered ‘No’ to at least one of the following questions (and ‘Yes’ to none) were defined as controls: ‘Have you ever been diagnosed with cancer, including skin cancer or cancerous moles?’, ‘Has a doctor ever told you that you have a type of cancer?’, ‘Have you ever been diagnosed or treated with any of the following conditions?’ Among the samples with imputed genotypes, 23andMe has 12,945 BCC cases and 274,252 controls.

Stage 2 study design and population. The Nurses’ Health Study was established in 1976, when 121,700 female registered nurses between the ages of 30 and 55 years residing in 11 larger US states completed and returned an initial self-administered questionnaire on their medical histories and baseline health-related exposures. Biennial questionnaires with collection of exposure information on risk factors have been collected prospectively. Roughly 20% of the genome was genotyped GWAS results that were only genotyped in the ‘V1’ platform were flagged due to small sample size, and SNPs on chrM or chrY because many of these are not currently called reliably. Using trio data, SNPs with a call rate $<0.6$ and $P < 10^{-20}$ for a test of $β = 1$ were flagged, SNPs with a Hardy-Weinberg $P < 10^{-10}$ in Europeans, or a call rate of $<90$%, were also flagged. Genotyped SNPs were also tested for genotype date effects, and SNPs with $P < 10^{-50}$ by analysis of variance of SNP genotypes against a factor dividing genotyping date into 20 roughly equal-sized buckets were flagged.

For the remaining SNPs, $P < 0.05$ or min.$\text{rsq}<0.3$ in any imputation batch were flagged, as well as SNPs that had strong evidence of an imputation batch effect. The batch effect test was an F test from an analysis of variance of the SNP dosage against a factor representing imputation batch; results with $P < 10^{-50}$ were flagged. Prior to GWAS, the largest subset of the data passing these criteria was identified for each SNP, based on their original genotyping platform—either v2 $\times$ v3 + v4, v3 $\times$ v4, v3 or v4 only—and association test results were computed for whatever was the largest passing set. As a result, there were no imputed results for SNPs that failed these filters. When choosing between imputed and genotyped GWAS results, if either the imputed or genotyped quality control, or a genotyping test was unavailable, the imputed result was reported; otherwise, the genotyped result was reported. For tests using imputed data, imputed dosages were used rather than best-guess genotypes.

Across all results, logistic regression results that did not converge due to complete separation, identified by abs (effect) $>10$ or stder $>10$ on the log odds scale, were not included. The Affymetrix, Illumina, and AffyOmni Express data were also flagged, since tests of low frequency variants can be sensitive to violations of the regression assumption of normally distributed residuals\(^20,54,55\).
heart disease). The Illumina HumanHap data set was comprised of several platforms: Illumina 550k (NHS-breast cancer, NHS-Pancreas cancer, HPFS-pancreas cancer, HPFS-kidney stone, HPFS-prostate cancer) and Illumina 660 (NHS-glaucoma, HPFS-glaucoma). The Illumina Omni Express data set contained only studies genotyped on the Omni Express platform (NHS-endometrial cancer, NHS-colon cancer, NHS-mammographic density, NHS-gout, HPFS-colon, HPFS-gout). We combined the individual data sets that were genotyped on the same platform, removing any SNPs that were not in all studies and with a missing call rate >5%, and flipping strands where appropriate to create a final compiled data set. This resulted in 688,283 SNPs in the Affymetrix data set, 459,999 SNPs in the Illumina HumanHap data set and 565,810 SNPs in the Illumina Omni Express data set. Analyses were restricted to subjects with self-reported European ancestry.

Genetic principal components were calculated using sets of independent SNPs (12,000–33,000 SNPs depending on platform). Subjects who did not cluster with other self-identified Europeans based on the top five principal components were also excluded.

We then ran a pairwise IBD analysis for each combined data set to detect duplicate and related individuals based on resulting Z scores. If 0 < Z ≤ 0.1 and 0.9 ≤ Z < 1.1 then a pair was flagged as being identical twins or duplicates. Pairs were considered full siblings if 0.17 ≤ Z ≤ 0.33 and 0.43 ≤ Z ≤ 0.6 and 0.17 ≤ Z < 0.33. Half siblings or avunculars were defined as having 0.4 ≤ Z ≤ 0.6 and 0 ≤ Z ≤ 0.1. Some of the duplicates flagged in this step were expected, having been genotyped in multiple data sets and hence having the same cohort IDs. In this case, one of each pair was randomly chosen for removal from the data set. Instances where pairs were flagged as unexpected duplicates with the different cohort IDs, but pairwise genotype concordance rate >0.99, resulted in removal of both individuals from the pair. Related individuals (full sibs, half sibs/avunculars) were not removed from the final data sets. In the Affymetrix data set, 167 individuals were removed because they were duplicates or were flagged for removal from secondary genotype data cleaning, leaving a total of 8063 individuals. Of the 6894 individuals originally in the Illumina data set, 107 were removed because they were duplicates or flagged for removal in the genotyping step, leaving 6787 IDs. In addition, eight pairs of individuals were flagged as related. In the Omni express data set, there were 5956 individuals at the start, with 39 IDs to remove leaving 5917 IDs and 5 pairs of related IDs.

After removing duplicate IDs and flagged related pairs of IDs, we used eigenstrat to run PCA analysis on each compiled data set, removing one member from each flagged pair of related individuals. For Affymetrix and Illumina HumanHap, we used approximately 12,000 SNPs that were filtered to ensure low pairwise LD9. For the OmniExpress data set we used approximately 33,000 SNPs that were similarly filtered. We plotted the top eigenvectors using R and examined the plots for outliers.

Finally as a quality control check, we ran logistic regression analyses using each individual study’s controls as ‘cases’ and the rest of the studies controls as ‘controls’. For example, in the Illumina Omni Express data set, we ran regressions of NHS-gout controls considered as ‘cases’ versus the HPFS-gout, NHS-endoctal cancer, NHS-colon cancer, NHS-mammographic density and HPFS-colon cancer. We then ran regressions with each of the other study controls as ‘cases’ versus all of the rest of the controls. We looked for P values of genome-wide significance (P < 5 × 10−8) and examined QQ plots to determine if any SNPs were flagged as significant where no SNPs should have been significant. In the Affymetrix data set 100 SNPs were flagged and removed. In the Illumina HumanHap data set, eight SNPs had P < 10−8 in any of the QC regressions and were removed. No SNPs in the Illumina Omni Express data set had P-values ≤ 8 × 10−8 hence no additional SNPs needed to be removed. After the data sets were combined and appropriate SNP and ID filters applied, the compiled data sets were imputed.

Using combined GWAS genotypes on each genotyping platform and the 1000 Genomes Project ALL Phase I Integrated Release Version 3 Haplotypes (excluding monomorphic and singleton sites (2010-11 data freeze, 2012-03-14 haplotypes)) as reference panel, we imputed the genotypes of markers in the 1000 Genomes Project for 8065 samples in Affymetrix data set, 6787 samples in Illumina HumanHap data set and 5917 samples in Illumina Omni Express data set.

SNP genotypes were imputed in three steps. First, genotypes on each chromosome were split into chunks to facilitate windowed imputation in parallel using ChunkChromosome (v.2011-08-05) (http://genome.sph.umich.edu/wiki/ChunkChromosome). Then each chunk of chromosome was phased using MACH (v.1.0.18c) (http://www.sph.umich.edu/csg/abecasis/MaCH/index.html). In the final step, Minimac (v.2012-08-15) (http://genome.sph.umich.edu/wiki/Minimac) was used to impute the phased genotypes to approximately 31 million markers in the 1000 Genomes Project11.

Sensitivity analysis in stage 2. Sensitivity analysis for five SNPs—rs12203592, KRT5, rs9267650, PLIN3, rs1408799 and EXOC2—using high imputation quality imputation was performed. Co-ancestry for two SNPs (rs12916300 and rs35407) was compared between stage 2 imputed data and directly genotyped data among the overlapping samples of 251 controls and 280 cases. Direct genotyping was also performed for rs9275642 and rs1050529 in a subset of BCC cases and controls within NHS. The total sample size is 1204 with 661 cases and 543 controls.

Stage 2 phenotype categorization. Participants in both NHS and HPFS cohorts reported new BCC diagnosis biennially. Eligible cases in the NHS and HPFS consisted of participants with self-reported BCC any time after baseline up to the 2012 follow-up cycle for both cohorts. Samples free of BCC were controls in this study. In the three compiled data sets, samples without information on BCC diagnosis were excluded. Furthermore, all BCC cases with melanoma were also excluded. Among the samples with imputed genotypes, we had 1777 BCC cases and 5411 controls in Affymetrix data set, 1268 BCC cases and 3685 controls in Illumina HumanHap data set and 1197 BCC cases and 3706 controls in Illumina Omni Express data set totalling 4242 BCC cases and 12,802 controls.

Validity of stage 2 self-reported data. The identification of BCC cases in stage 2 was based on self-report without pathological confirmation. Because the participants in the cohorts were nurses and other health professionals, the validity of their reports was expected to be high and has been proven in validation studies: >90% confirmed by histopathology records57–59. In addition, previous studies of BCC in the NHS using self-reported cases identified both constitutional and sun-exposure risk factors as expected, such as lighter pigmentation, less childhood and adolescent tanning tendency, higher tendency to sunburn, and tanning salon attendance58,60. Moreover, in our previous study of BCC using the same cohorts as this study (NHS and HPFS), we confirmed the MCH1 gene (a well-established pigmentation gene) as the most promising locus of BCC risk5. In addition, we identified many other pigmentation related SNPs (for example, rs12203592 at IRF4, rs1408799 at TYRP1 and rs12913832 at HERC2) in our data set with significant associations with BCC4. These genetic and non-genetic data together suggest that the bias due to self-report of BCC is likely to be minimal in our study.

Stage 2 genome-wide association analysis. We used ProbABEL software to test the GWAS association between minor allele counts and BCC risk using imputed dosage data. We performed logistic regression analysis under an additive model with adjustment for age, sex, BCC history and the first five principal components, generating the following model:

BCC diagnosis ~ age + sex + pc1 + pc2 + pc3 + pc4 + pc5 + genotype (2)

These principal components were calculated for all individuals on the basis of approximately 10,000 unlinked markers using the EIGENSTRAT software6. Associations in each component GWAS set (Affymetrix, Illumina HumanHap series and Illumina Omni Express) were combined in an inverse-variance-weighted meta-analysis using the METAL software.

Phenotype and environmental factors. The amount of UV exposure based on residence location in 1986 was coded as a continuous variable. Information on self-reported sun-protective behaviors (for example, use of a sunscreen, wearing a hat, using sunglasses) was collected during the 1986–1989 baseline survey in the NHS and HPFS using similar wording. We regressed ordinal coding for natural hair colour (1—red, 2—blonde, 3—light brown, 4—dark brown or black), tanning ability during adolescence (1—tans without burning (none/some redness only), 2—burns then tans, 3—burns/peels (painful burn with blisters)) and number of blistering sunburns (1 = never, 2 = 1–2 times, 3 = 3–5 times, 4 = 6 times and above).

Statistical models for pigmentation and UV exposure. To clarify the influence of pigmentation traits on our associations, we additionally adjusted for these traits (hair colour, tanning ability during adolescence and number of blistering sunburns) in individual studies with ProbABEL. Associations in each component GWAS set (Affymetrix, Illumina HumanHap series and Illumina Omni Express) were combined in the same meta-analysis using the METAL software.

The primary gene-risk factor interaction analytic model included each SNP and risk factor (UV exposure or pigmentation traits) and an SNP × risk factor interaction term, as well as age, sex, and the first five principal components as covariates:

BCC diagnosis ~ age + sex + risk factor + PCA1 + PCA2 + PCA3 + PCA4 + PCA5 + genotype + genotype × risk factor (3)

The logistic regression analyses within each platform were conducted with ProbABEL. Associations in each component GWAS set (Affymetrix, Illumina HumanHap series and Illumina Omni Express) were combined in an inverse-variance-weighted meta-analysis using the METAL software.

Meta-analysis. For each SNP, meta-analysis was conducted to combine stage 1 and stage 2 results. Heterogeneity of per-SNP effect sizes in studies contributing to stage 1, stage 2 and the overall meta-analysis was assessed and fixed effects meta-
Proportion of familial relative risk. We have used the formula for calculating the proportion of FRR as outlined by the Cancer Oncological Environment Study (http://www.nature.com/icogi/primer/common-variation-and-heritability-estimates-breast-ovarian-prostate-cancers/70) as described previously.63 The odds ratios derived from our meta-analysis of stages 1 and 2 are assumed to be relative risks. We estimated the proportion of the FRR explained by each SNP (FRRexp) as:

\[
\text{FRR}_{\text{exp}} = \left( \frac{p^2 + q^2}{p^2 + q^2} \right)^2
\]

(4)

Here, the risk allele and alternative allele frequencies are p and q, respectively, and \( r \) is the odds ratio for the risk allele. Allele frequencies derived from the 1000 Genomes European population data. Assuming that the loci combine multiplicatively and are not in LD, the combined effect of all loci is given by:

\[
\lambda_r = \prod_k \lambda_k
\]

(5)

Here, the product is across all loci. The proportion of the familial relative risk attributable to the SNPs on a log scale, is then given by:

\[
\log \left( \frac{\lambda_r}{\lambda_p} \right) = \log \left( \frac{\lambda_p}{\lambda_r} \right)
\]

(6)

In this equation, \( \lambda_p \) is the familial relative risk observed in epidemiological studies. \( \lambda_p \) is 2.6-fold for BCC.64

Regulatory function of novel variants. For each novel BCC susceptibility variant, we searched for evidence of regulatory function using recently updated HaploReg version 4 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php).65 We queried each rsID and extracted data from ENCODE Project Consortium 2011–2012 on closest annotated gene, ChIP-Seq transcription factor binding, DNaseI hypersensitivity sites, and enhancer and promoter chromatin segmentation states66–69. We also extracted data from Roadmap Epigenomics Consortium 2015 on enhancer and promoter chromatin segmentation states, specifically using the following states: 15-state HMM, 25-state HMM, H3K4me1, H3K4me3, H3K27ac and H3K36ac.70 We particularly focused on enhancer and promoter annotations that referenced normal human epithelial keratinocytes (NHKEK) and primary foreskin keratinocytes. Finally, we used HaploReg v4 to extract eQTL data for each variant, as version 4 is updated with cis eQTL data from the GTEx pilot analysis and other studies.11 We made special note of variants that were eQTLs in skin tissue.

Gene expression analysis. Processed gene-expression data for BCC and normal skin (GSE53462 and GSE5753) were obtained from the GEO (http://www.ncbi.nlm.nih.gov/geo).71 Fifteen BCC samples and five controls were included for GSE5753; for GSE53462, nine ‘classic-type’ BCC controls were included for GSE7553; for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classic-type’ BCC controls were included for GSE53462, nine ‘classi...
45. Lessard, L.

28. Ren, L.-R.

31. Tillotson, L. G. RIN ZF, a novel zinc finger gene, encodes proteins that bind to

expression is controlled by the transcription factors Sp1 and Sp3.

631–638 (2009).

is essential for Toll-like receptor activation and cellular signaling.

16, 16, Genet.

1735–1742 (2015).

5750–5762 (2014).

1084–1097 (2007).

81, 81, 8123–8128

776–784 (2013).

955–965 (2007).

7851–7865

63. Law, M. H.

66. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, and

56. Yu, K.

411 (2000).

1514–1521 (2006).

Price, A. L. et al. Principal components analysis corrects for stratification in

135, 659–667

J. Immunol.

776–781 (2015).

75. Pruim, R. J.

77. Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J. L. Trend tests for case-control
disease outcomes in a prospective cohort study of women.

2002–2006 (2014).

12, 135, 569–667 (2010).

Pan, X. et al. The transcription factor CUT1 is associated with proliferation and survival in malignant melanoma. Melanoma Res. 24, 198–206 (2014).

Al Olama, A. A. et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet. 46, 1103–1109 (2014).

Qian, X. et al. The Tensin-3 protein, including its SH2 domain, is phosphorylated by Src and contributes to tumorigenesis and metastasis. Cancer Cell 16, 246–258 (2009).

Chen, X. et al. RNASE12 tag SNP but not CCR6 polymorphisms is associated with autoimmune thyroid diseases in the Chinese Han population. BMC Med. Genet. 16, 11 (2015).

Lualdi, M. et al. Pleiotropic modes of action in tumor cells of RNASE12, an evolutionary highly conserved extracellular RNase. Oncotarget 6, 7851–7865 (2015).

Monti, L. et al. RNASE12 is a tumor antagonizing gene in a melanoma cancer model. Oncol. Res. 17, 69–74 (2008).

Wang, Q. et al. Stress-induced RNASE12 overexpression mediates melanocyte apoptosis via the TRAF2 pathway in vitro. Cell Death Dis. 5, e1022 (2014).

Sand, M. et al. Long non-coding RNAs in basal cell carcinoma. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med. 37, 1–14 (2016).

Lessard, L. et al. The CASC15 long intergenic noncoding RNA locus is involved in melanoma progression and phenotype switching. J. Invest. Dermatol. 135, 2464–2474 (2015).

Visser, M., Palstra, R.-J. & Kayser, M. Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. Hum. Mol. Genet. 23, 5750–5762 (2014).

Jaconi, L. C. et al. A genome wide association study identifies the skin color genes IRF4, MC1R, ASIP, and BNC2 influencing facial pigmented spots. J. Invest. Dermatol. 135, 1735–1742 (2015).

Eriksson, N. et al. Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet. 6, e1000993 (2010).

Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. Ancestry Composition: A Simple Efficient Pipeline for Ancestry Decovolution (2014).

Henn, B. M. et al. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 7, e43267 (2012).

1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).

Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).

53. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. Bioinforma. Oxf. Engl. 31, 782–784 (2015).

Jorgensen, E. et al. A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. Nat. Commun. 6, 10130 (2015).

Hu, Y. et al. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nat. Commun. 7, 10448 (2016).

Yu, K. et al. Population substructure and control selection in genome-wide association studies. PLoS One 3, e2551 (2008).

Colditz, G. A. et al. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. Am. J. Epidemiol. 123, 894–900 (1986).

Hunter, D. J. et al. Risk factors for basal cell carcinoma in a prospective cohort of women. Ann. Epidemiol. 1, 13–23 (1990).

van Dam, R. M. et al. Risk factors for basal cell carcinoma of the skin in men: results from the health professionals follow-up study. Am. J. Epidemiol. 150, 447–458 (1999).

Han, J., Colditz, G. A. & Hunter, D. J. Risk factors for skin cancers: a nested case-control study within the Nurses’ Health Study. Int. J. Epidemiol. 35, 1514–1521 (2006).

Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909 (2006).

Johnson, A. D. et al. SNP a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinforma. Oxf. Engl. 24, 2938–2939 (2008).

Law, M. H. et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. Nat. Genet. 47, 987–995 (2015).

Melnik, N. L. et al. Family history of skin cancer is associated with early-onset basal cell carcinoma independent of MCIR genotype. Cancer Epidemiol. 39, 1078–1083 (2015).

Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 22, 1798–1812 (2012).

Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40, D930–D934 (2012).

Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).

ENCODE Project Consortium. A user’s guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9, e1001046 (2011).

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

Chadwick, L. H. The NIH Roadmap Epigenomics Program data resource. Epigenomics 4, 317–324 (2012).

Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210 (2002).

Lee, B. A. et al. Molecular classification of basal cell carcinoma of skin by gene expression profiling. Mol. Carcinog. 54, 1605–1612 (2015).

Riker, A. I. et al. The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. BMC Med. Genomics 1, 13 (2008).

Davia, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinforma. Oxf. Engl. 23, 1846–1847 (2007).

Prunin, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinforma. Oxf. Engl. 26, 2336–2337 (2010).

Gordon, M. Package forestplot’. Advanced Forest Plot Using grid Graphics (2016).

Freidlin, B., Zheng, G., Li, Z. & Gastwirth, J. L. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum. Hered. 53, 146–152 (2002).

Acknowledgements

We would like to thank the research participants and employees of 23andMe for making this work possible. Additionally, we would like to thank Constance Chen for her technical support and the participants and staff of the Nurses’ Health Study, the Health Professionals Follow-Up Study for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NI, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. This work was supported by the National Human Genome Research Institute of the National Institutes of Health [grant number R44HG006981], and in part by NIH R01 CA49449, P01 CA86157, U1. CA167552; Stanford Medical Scholars Research Program (H.C.), Stanford TRAM (Translational Research and Applied Medicine program), and the Dermatology Foundation Career Development Award (K.S.).
Author contributions
J.H., K.S. and J.T. designed and oversaw the study. W.W., H.C., J.H., D.H. and K.S. were responsible for quality control, manuscript writing and data analyses. J.H., D.H., P.K., W.L. and A.Q. contributed to data acquisition. K.R., J.T., K.S. and H.C. carried out survey validation collection and analysis. Bioinformatics analyses were carried out by W.W., H.C., D.H., Y.L., H.D., J.H. and K.S. Especially major contributions to writing and editing were made by W.W., H.C., J.H., K.S., D.H. and J.T. All authors contributed to and critically reviewed the manuscript.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: David Hinds is an employee at 23andMe. The remaining authors declare no competing financial interests.

Reprints and permission information is available at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Chahal, H. S. et al. Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. Nat. Commun. 7:12510 doi: 10.1038/ncomms12510 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third-party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016