Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma

Citation

Published Version
doi:10.1038/ncomms12048

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:29407815

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma

Harvind S. Chahal1,*, Yuan Lin2,*, Katherine J. Ransohoff1, David A. Hinds3, Wenting Wu2, Hong-Ji Dai2,4, Abrar A. Qureshi5,6,7, Wen-Qing Li5,6, Peter Kraft8,9, Jean Y. Tang1, Jiali Han2,4,8,** & Kavita Y. Sarin1,**

Cutaneous squamous cell carcinoma represents the second most common cutaneous malignancy, affecting 7–11% of Caucasians in the United States. The genetic determinants of susceptibility to cutaneous squamous cell carcinoma remain largely unknown. Here we report the results of a two-stage genome-wide association study of cutaneous squamous cell carcinoma, totalling 7,404 cases and 292,076 controls. Eleven loci reached genome-wide significance ($P < 5 \times 10^{-8}$) including seven previously confirmed pigmentation-related loci: \(MC1R\), \(ASIP\), \(TYR\), \(SLC45A2\), \(OCA2\), \(IRF4\) and \(BNC2\). We identify an additional four susceptibility loci: 11q23.3 \(CADM1\), a metastasis suppressor gene involved in modifying tumour interaction with cell-mediated immunity; 2p22.3; 7p21.1 \(AHR\), the dioxin receptor involved in anti-apoptotic pathways and melanoma progression; and 9q34.3 \(SEC16A\), a putative oncogene with roles in secretion and cellular proliferation. These susceptibility loci provide deeper insight into the pathogenesis of squamous cell carcinoma.
Squamous cell carcinoma (SCC) of the skin represents the second most common cutaneous malignancy, behind only basal cell carcinoma (BCC), with a lifetime incidence of 7–11% in Caucasians in the United States, creating a substantial economic and health burden. In addition, an estimated 6,000 deaths are attributed to SCC annually in the United States. Fair skin, male gender, ultraviolet radiation exposure and prior organ transplant are associated with increased rates of SCC.

In contrast to BCC, few genetic variants have thus far been linked to SCC risk. A recent genome-wide association study (GWAS) from patients within the Kaiser Permanente Healthcare System identified ten SCC susceptibility loci, including six pigmentation loci: SLC45A2, IRF4, TYR, HERC2, DEF8 (16q24, the same locus as MC1R) and RALY2. Here we report a two-stage genome-wide association meta-analysis for SCC, with a total of 7,404 SCC cases and 292,076 controls from the 23andMe research participant cohort and the Nurses’ Health Study/Health Professionals Follow-Up Study (HPFS). Our results provide independent replication for nine of the ten loci from the Kaiser Permanente GWAS and identify four additional novel susceptibility loci for SCC.

Results

**Genome-wide association study.** Stage 1 consisted of a GWAS set encompassing 6,579 self-reported SCC cases and 280,558 controls of European ancestry from the 23andMe research participant cohort (Table 1). A validation study of the 23andMe self-reported data was conducted, which revealed a sensitivity and specificity of 92% and 98%, respectively (Supplementary Table 1). Self-reported data was conducted, which revealed a sensitivity and specificity of 92% and 98%, respectively (Supplementary Table 1). Stage 2 consisted of an independent GWAS set of 825 adjudicated SCC cases and 11,518 controls of European ancestry from the Nurses’ Health Study and HPFS (Table 1) to confirm the index SNPs at each locus associated with SCC (\(P<10^{-7}\)) was identified, yielding a list of 14 index SNPs (Fig. 1). Stage 2 provided independent replication for nine of the ten loci from the Kaiser Permanente GWAS and identify four additional novel susceptibility loci for SCC.

Confirmation of previously reported loci. Among the 11 genome-wide significant loci identified in this two-stage study (Table 2), 7 were previously reported in the Kaiser GWAS but had not yet been replicated in an external cohort (Table 3, Supplementary Table 2 and Supplementary Fig. 9). Of the seven previously reported loci, six are pigmentation related, whereas the seventh, 9p22.2 BNC2, is a putative pigmentation locus3. In addition to replicating these seven loci at genome-wide significance, our study replicated two additional loci from the Kaiser GWAS, 3p13 FOXP1 and 3q28 TPRG1/TP63 (\(P = 4.4 \times 10^{-3}\) and \(P = 1.3 \times 10^{-5}\), respectively, by logistic regression) (Table 3 and Supplementary Table 2).

**Genome-wide significant novel susceptibility loci.** Our study also identified four novel SCC susceptibility loci (Table 2, Supplementary Table 3 and Supplementary Fig. 10) not previously reported. These four novel susceptibility loci—2p22.3, 7p21.1 (AHR), 9q34.3 (SEC16A) and 11q33.3 (CADM1-BUD13)—have not been previously associated with pigmentation phenotypes. Of these four loci, one (SEC16A) reached genome-wide significance in stage 1, whereas the other three did so in the combined meta-analysis. Although some loci did not reach statistical significance in stage 2, their 95% confidence intervals (for odds ratios (ORs)) overlapped with corresponding stage 1 confidence intervals. The power to reach \(P<0.05\) in stage 2 for all 14 index SNPs is detailed in Supplementary Table 4.

### Table 1 | Gender and age of cases and controls from each stage of GWAS.

<table>
<thead>
<tr>
<th>Status</th>
<th>n (%)</th>
<th>Male (%)</th>
<th>Age &lt; 31 years</th>
<th>Age 30–45 years</th>
<th>Age 46–60 years</th>
<th>Age &gt; 60 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>23andMe (Stage 1, (n = 287,137))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>6,579 (2.3)</td>
<td>3,510 (53)</td>
<td>12 (0.18)</td>
<td>199 (3)</td>
<td>1,263 (19)</td>
<td>5,105 (78)</td>
</tr>
<tr>
<td>Controls</td>
<td>280,558 (97.7)</td>
<td>151,588 (54)</td>
<td>39,838 (14)</td>
<td>83,780 (30)</td>
<td>76,833 (27)</td>
<td>80,107 (28)</td>
</tr>
<tr>
<td>Harvard (Stage 2, (n = 12,343))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>367 (6.3)</td>
<td>183 (49.9)</td>
<td>0 (0.0)</td>
<td>91 (24.8)</td>
<td>213 (58.0)</td>
<td>63 (17.2)</td>
</tr>
<tr>
<td>Controls</td>
<td>5,453 (93.7)</td>
<td>2,412 (44.2)</td>
<td>0 (0.0)</td>
<td>1,952 (36.8)</td>
<td>2,729 (50.0)</td>
<td>772 (14.2)</td>
</tr>
<tr>
<td>Illumina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>220 (7.0)</td>
<td>91 (41.4)</td>
<td>0 (0.0)</td>
<td>61 (27.7)</td>
<td>125 (56.8)</td>
<td>34 (15.5)</td>
</tr>
<tr>
<td>Controls</td>
<td>2,901 (93.0)</td>
<td>232 (8.0)</td>
<td>0 (0.0)</td>
<td>1,366 (47.1)</td>
<td>1,434 (49.4)</td>
<td>101 (3.5)</td>
</tr>
<tr>
<td>Omni</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>238 (7.0)</td>
<td>102 (42.9)</td>
<td>0 (0.0)</td>
<td>72 (30.2)</td>
<td>137 (57.6)</td>
<td>29 (12.2)</td>
</tr>
<tr>
<td>Controls</td>
<td>3,164 (93.0)</td>
<td>803 (25.4)</td>
<td>0 (0.0)</td>
<td>1,401 (44.3)</td>
<td>1,488 (47.0)</td>
<td>275 (8.7)</td>
</tr>
<tr>
<td>All, Stage 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>825 (6.7)</td>
<td>376 (45.6)</td>
<td>0 (0.0)</td>
<td>224 (27.1)</td>
<td>475 (57.6)</td>
<td>126 (15.3)</td>
</tr>
<tr>
<td>Controls</td>
<td>11,518 (93.3)</td>
<td>3,447 (29.9)</td>
<td>0 (0.0)</td>
<td>4719 (41.0)</td>
<td>5651 (49.1)</td>
<td>1148 (9.9)</td>
</tr>
<tr>
<td>Combined meta-analysis ((n = 29,9480))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases</td>
<td>7,404 (2.5)</td>
<td>3,886 (52)</td>
<td>12 (0.16)</td>
<td>423 (5.7)</td>
<td>1,738 (23)</td>
<td>5,231 (71)</td>
</tr>
<tr>
<td>Controls</td>
<td>29,2076 (97.5)</td>
<td>155,035 (53)</td>
<td>39,838 (14)</td>
<td>88,499 (30)</td>
<td>82,484 (28)</td>
<td>81,255 (28)</td>
</tr>
</tbody>
</table>

GWAS, genome-wide association study.
Counts and percentages for cases and controls (\(n\%)) are listed above, stratified by stage of GWAS. We also report number and percentage of male subjects, subjects with age < 31 years, subjects with age 31–45 years, subjects with age 46–60 years and subjects with age > 60 years. Stage 2 cases and controls are further subdivided based on platform used for genotyping.
Heritability of SCC. To measure the proportion of SCC heritability that can be attributed to these SNPs, we calculated the familial relative risk for SCC as outlined by the Cancer Oncological Gene–Environment Study. Overall, 6.3% of familial relative risk for SCC is explained by these 11 loci. Interestingly, although consistent effects for all 7 SNPs were observed across age and gender, larger effect sizes tended to occur in younger cases, perhaps highlighting the increasingly important influence of environmental factors with age (Supplementary Tables 5 and 6, and Supplementary Fig. 11). Further results can be found in Supplementary Tables 7–11 and Supplementary Fig. 12.

Discussion

Previously reported loci. The predominance of SCC susceptibility loci associated with pigmentation genes confirms the well-established heritable phenotypic risk of SCC. In our study, these SNPs include the following: rs1805007 (MCIR R151C), a red hair allele associated with photosensitivity and increased BCC risk; rs12203592, which lies within an enhancer of IRF4 transcription in melanocytes and is associated with increased risk of actinic keratoses (SCC precursors) independent of skin pigmentation; rs1126809 (TYR R402Q), associated with photosensitivity, tanning and increased risk of BCC and melanoma; rs6059655, intergenic near RALY-ASIP.
and associated with facial pigmented spots\(^{10}\); and rs35407, in modest linkage disequilibrium with rs16891982 (SLC45A2 F374L, \(r^2 = 0.33, D^2 = 1\)), associated with human pigmentation and melanoma risk\(^{11,12}\). rs16891982 reached genome-wide significance in stage 1 (\(P = 8.6 \times 10^{-11}\), OR = 1.65, logistic regression) and in the overall meta-analysis (\(P = 1.5 \times 10^{-11}\), OR = 1.62, logistic regression) (Supplementary Table 7). We also performed sensitivity analysis on OCA2 rs1800407 and IRF4 rs12203592 using high imputation quality subsets and directly genotyped data sets (Supplementary Table 8)\(^{13}\).

In addition to these confirmed polymorphism loci, rs10810657 at 9p22.2 reached genome-wide significance in the overall meta-analysis (\(P = 1.4 \times 10^{-8}\), OR = 0.90, logistic regression). rs10810657 lies \(~13\) kb upstream of the basonuclin 2 (BN2) transcription start site in a potential enhancer in foreskin melanocytes and other cell types\(^{8}\). This SNP is in linkage disequilibrium with rs12350739 (\(r^2 = 0.90, D^2 = 1.00\)), which resides in an enhancer element regulating BN2 transcription in human melanocytes\(^{14}\). BN2 is a DNA-binding zinc-finger protein thought to act as both a messenger RNA-processing enzyme and a transcription factor\(^{14}\). BN2 is expressed in melanocytes and, to a lesser extent, keratinocytes, with higher expression levels corresponding to darker skin pigmentation in human skin tissue variants. Variants in the BN2 locus have recently been associated with skin colour, freckling and age-related pigmentation spots in Europeans, in addition to SCC\(^{3,10,15}\). The association and linkage disequilibrium results for these SNPs are listed in Supplementary Tables 9 and 10.

### Table 3 | Replication of ten previously reported SCC-associated loci.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Region</th>
<th>Gene</th>
<th>Major/minor</th>
<th>MAF (avg imp (r^2))</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Meta-analysis</th>
<th>Prior GWAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12203592</td>
<td>6p25.3</td>
<td>IRF4</td>
<td>C/T</td>
<td>0.17 (0.99)</td>
<td>3.0 × 10^{-10}</td>
<td>1.62</td>
<td>2.1 × 10^{-5}</td>
<td>1.56</td>
</tr>
<tr>
<td>rs1805007</td>
<td>1q62.4</td>
<td>MCIR</td>
<td>C/T</td>
<td>0.07 (1.0)</td>
<td>5.8 × 10^{-3}</td>
<td>1.45</td>
<td>9.5 × 10^{-5}</td>
<td>1.64</td>
</tr>
<tr>
<td>rs6059655</td>
<td>20q11.22</td>
<td>RALY-ASIP</td>
<td>G/A</td>
<td>0.07 (0.99)</td>
<td>3.6 × 10^{-14}</td>
<td>1.28</td>
<td>8.2 × 10^{-1}</td>
<td>1.03</td>
</tr>
<tr>
<td>rs35407</td>
<td>5p13.3</td>
<td>SLC45A2</td>
<td>G/A</td>
<td>0.04 (0.98)</td>
<td>3.6 × 10^{-15}</td>
<td>0.59</td>
<td>3.6 × 10^{2}</td>
<td>0.58</td>
</tr>
<tr>
<td>rs1126809</td>
<td>1q14.3</td>
<td>TYR</td>
<td>G/A</td>
<td>0.28 (0.99)</td>
<td>3.5 × 10^{-14}</td>
<td>1.17</td>
<td>8.1 × 10^{-1}</td>
<td>1.02</td>
</tr>
<tr>
<td>rs1800407</td>
<td>15q13.1</td>
<td>OCA2</td>
<td>C/T</td>
<td>0.07 (1.0)</td>
<td>8.0 × 10^{-9}</td>
<td>1.21</td>
<td>8.9 × 10^{-9}</td>
<td>0.98</td>
</tr>
<tr>
<td>rs10810657</td>
<td>9p22.2</td>
<td>BN2, CNTN</td>
<td>A/T</td>
<td>0.41 (0.98)</td>
<td>2.4 × 10^{-7}</td>
<td>0.91</td>
<td>7.9 × 10^{-3}</td>
<td>0.82</td>
</tr>
<tr>
<td>rs6224607</td>
<td>1p31.3r</td>
<td>FOXP1</td>
<td>G/A</td>
<td>0.33 (0.85)</td>
<td>3.2 × 10^{-3}</td>
<td>1.06</td>
<td>9.6 × 10^{-3}</td>
<td>1.00</td>
</tr>
<tr>
<td>rs6791479</td>
<td>3q28</td>
<td>TPRG3/TP63</td>
<td>A/T</td>
<td>0.43 (0.99)</td>
<td>7.0 × 10^{-2}</td>
<td>1.03</td>
<td>3.5 × 10^{-3}</td>
<td>1.23</td>
</tr>
<tr>
<td>rs4455710</td>
<td>6p21.2</td>
<td>HLA-DQA1</td>
<td>C/T</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Avg imp, average imputation; GWAS, genome-wide association study; MAF, minor allele frequency; OR, odds ratio; SCC, squamous cell carcinoma; SNP, single-nucleotide polymorphism.

Ten loci previously reported as associated with SCC via prior GWAS (\(P < 5 \times 10^{-8}\)) are listed, the first seven of which also reached genome-wide significance in this study. Of the remaining three loci (asterisks), two (3p13 and 3q28) reached nominal significance (\(P < 0.05\)) in this study. In addition, we report genetic locus, nearest genes, major allele, minor allele, MAF as calculated from stage 1 data, unadjusted OR (a measure of imputation quality) for stage 1 OR with P-value for each stage, calculated with respect to the minor allele. The right-most two columns list P-value and OR from prior GWAS for each locus, relative to the minor allele.

**Novel susceptibility loci.** We also identified four novel SNPs associated with SCC. rs57994353 at 9q34.3 (\(P = 7.5 \times 10^{-9}\), OR = 1.12, logistic regression) resides within an intron of SEC61A in tight linkage disequilibrium with rs3812594 SEC16A R1039C (\(r^2 = 0.90, D^2 = 1\))\(^{8}\). SEC16A is a cytosolic scaffold protein that acts at endoplasmic reticulum exit sites to facilitate vesicle formation and export\(^{16}\). Production of this protein is regulated by various growth factors, with higher levels of SEC16A corresponding to increased secretion and cellular proliferation\(^{17}\). Overexpression of SEC16A has been observed in colonic adenocarcinoma samples\(^{17}\).

Another novel SCC risk variant, rs74899442 at 11q23.3, reached genome-wide significance in the overall meta-analysis (\(P = 8.7 \times 10^{-9}\), OR = 2.13, logistic regression). This SNP lies in an intergenic region 515 kb upstream of CADM1, which encodes a single-pass transmembrane protein involved in cell-cell adhesion and signal transduction\(^{18-20}\). CADM1 is also a putative tumour suppressor in many human carcinomas, including SCC, as it is frequently downregulated in these tissues via promoter methylation\(^{21-23}\). Notably, CADM1 expression levels are associated with survival in SCC patients. In a study of 87 patients with SCC, those with decreased CADM1 expression levels are associated with survival in SCC patients. We also identified four novel SNPs associated with SCC via prior GWAS (\(P < 5 \times 10^{-8}\)) in this study. In addition, we report genetic locus, nearest genes, major allele, minor allele, MAF as calculated from stage 1 data, unadjusted OR (a measure of imputation quality) for stage 1 OR with P-value for each stage, calculated with respect to the minor allele. The right-most two columns list P-value and OR from prior GWAS for each locus, relative to the minor allele.

**Suggestive novel susceptibility loci.** We also provide evidence of additional SCC susceptibility loci. Three loci, 6p21.32 (rs28993540), 3q28 (rs11715549) and 8q23.3 (rs199816436), with high imputation quality (\(R^2 > 0.9\)) were associated with SCC risk in stage 1 (\(P < 10^{-5}\)) but did not reach \(P < 0.05\) in stage 2, or \(P < 5 \times 10^{-8}\) in the overall meta-analysis (Supplementary Table 11 and Supplementary Fig. 12). rs28993540 lies 35 kb upstream of HLA-DQB1. Transplant and other immunosuppressed patients are at significantly increased risk of SCCs suggesting a role for T-cell-mediated protection against SCC\(^{27}\). HLA-DQB1 alleles have been associated with risk of squamous cell cervical cancer and may function by altering the efficiency of the T-cell-mediated immune response to HPV antigens\(^{28}\). rs11715549 resides in a potential
enhancer in keratinocytes within LPP, a gene involved in epithelial cell motility and cell–cell adhesion. rs11715549 is in tight linkage disequilibrium with SNPs associated with vitiligo (rs9851967, $r^2 = 0.93$), celiac disease (rs1464510, $r^2 = 0.74$) and allergy (rs9860547, $r^2 = 0.87$)$^{12,29}$. rs199816436 lies inside an intron in TRPS1, a transcription factor that represses GATA-regulated genes and binds to a dynem light chain protein. rs199816436 tightly correlated with rs10808475 ($r^2 = 0.85$) and is a TRPS1 expression quantitative trait locus (eQTL) in liver$^{30}$. Defects in TRPS1 lead to trichorhinophalangeal syndrome, a genetic syndrome characterized by coarse facies, brittle hair and skeletal defects. Although falling short of genome-wide significance, this evidence is nonetheless suggestive of an association between these loci and SCC that merits further investigation.

This two-stage meta-analysis provides the first independent replication of nine of ten previously reported SCC susceptibility loci and identifies four novel susceptibility loci. In addition, this large-scale GWAS demonstrates the power of consumer self-reported data from internet platforms as a resource for discovering cancer susceptibility loci, with results consistent with studies using adjudicated cancer data.

**Methods**

**Stage 1 study design and population.** 23andMe (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 take part in this research, in accord with 23andMe’s human subjects protocol (reviewed and approved by Ethical and Independent Review Services, an Association for the Accreditation of Human Research Protection Program (AAHRPP)-accredited Institutional Review Board (IRB)). 23andMe gathers genetic information by genotyping sample material provided by its customers; phenotypic information is collected via customer responses to online surveys. Inclusion and exclusion criteria are discussed below.

**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:

\[
\text{SCC diagnosis} \sim \text{age} + \text{sex} + \text{pc.0} + \text{pc.1} + \text{pc.2} + \text{pc.3} + \text{pc.4} + \text{genotype}.
\]

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. In addition, test statistics were adjusted for genotyping platform, to correct for any batch effect. This was implemented by controlling for principal components that failed to a test for parent-offspring transmission were also flagged; specifically, the child’s allele count was regressed against the mean parental allele count and SNPs with fitted $b < 0.6$ and $P < 10^{-15}$ for a test of $b = 1$ were flagged. SNPs with a Hardy–Weinberg $P < 10^{-20}$ in Ehrenstein’s European survey, 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 genotype data into 20 roughly equal-sized buckets were flagged.

For imputed GWAS results, SNPs with $P < 0.5$ or $r < 0.3$ in any imputation batch were flagged, as well as SNPs that had strong evidence of an imputation batch effect. The batch effect test was computed as an F-test for an analysis of variance of the SNP dosages against a factor representing imputation batch; results with $P < 10^{-50}$ were flagged. Before GWAS, the largest subset of the data passing these criteria was identified for each SNP, based on their original genotyping platform—either v2 + v3 + v4, v3 + v4 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 test passed quality control, or a genotyped 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 incomplete separation, identified by abs(effect) > 10 or stder > 10 on the log odds scale, were flagged. Linear regression results for SNPs with minor allele frequency <0.1% were also flagged, as tests of low-frequency variants can be sensitive to violations of the regression assumption of normally distributed residuals. This methodology has been applied in prior GWAS studies$^{36–38}$.

**Stage 1 phenotype categorization.** 23andMe identified SCC cases by using research participants’ self-reported answers to online questionnaires. Subjects who answered ‘Yes’ and/or selected SCC 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 squamous cell carcinoma?’ ‘What type of skin cancer did you have?’ ‘Have you ever been diagnosed with squamous 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 squamous 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 SCC 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?’ ‘Are you ever been diagnosed or treated with any of the following conditions?’ Among the samples with imputed genotypes, 23andMe has 6,579 SCC cases and 280,558 controls.

**Sensitivity and specificity of stage 1 self-reported data.** To assess the validity of self-reported phenotypic data in stage 1, 23andMe surveys (pertaining to skin cancer history and pigmentation) were randomly administered to 188 patients seen in Stanford outpatient clinics. The survey answers were then compared with medical records to assess for accuracy with respect to SCC diagnosis, to determine the sensitivity and specificity of the self-reported data. Values were determined using $\chi^2$-analysis. This sub-study was approved by the Stanford University Institutional Review Board with a waiver of documentation of consent.
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. Every 2 years, along with exposures, outcome data with appropriate follow-up of reported disease events are collected. Overall, follow-up has been high; after more than 20 years, ~90% of participants continue to complete questionnaires. From May 1989 through September 1990, we collected blood samples from 32,826 participants in the NHS. Information on SCC development was first collected in the 1984 questionnaire. The HPFS was established in 1986 when a cohort of men from all 50 US states in health professions (dentists, pharmacists, optometrists, osteopath physicians, podiatrists and veterinarians) aged 40–75 years answered a detailed mailed questionnaire. The average follow-up rate for this cohort over 10 years is >90%. On each biennial questionnaire, we obtained disease- and health-related information. Between 1993 and 1994, 18,159 study participants provided blood samples by overnight courier. Information on SCC development was first collected in the 1986 questionnaire.

The protocol for this study was approved by the Institutional Review Board at Brigham and Women’s Hospital and the Harvard School of Public Health. All of the participants provided informed consent.

Stage 2 genotyping and quality control. There were 18 GWAS data sets from the NHS and HPFS as nested case–control studies with cleaned genotypic data available. We combined these data sets into three compiled data sets based on their genotyping platform type: Affymetrix, Illumina HumanHap series or Illumina Omni Express. The Affymetrix data set comprised data on the Affy 610Q platform (NHS-type 2 diabetes, NHS-coronary heart disease, HPFS-type 2 diabetes and HPFS-coronary heart disease). The Illumina HumanHap data set comprises several platforms: Illumina 550 K (NHS-breast cancer, NHS-Pancreas cancer and HPFS-pancreas cancer), Illumina 610Q (NHS-kidney stone, HPFS-kidney stone and HPFS-prostate cancer) and Illumina 660 (NHS-glaucoma and 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 and HPFS-gout).

We combined the individual data sets that were genotyped on the same platform, removed any SNPs in all studies and with a missing call rate of ≥5%, and flipping strands, where appropriate, to create a final compiled data set. This resulted in 668,283 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 SNPs, 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 ≤ Z0 ≤ 0.1, 0.1 < Z1 ≤ 0.9 and 0.9 < Z2 ≤ 1.1, individuals were flagged as being identical twins or duplicates. Individuals or twins were flagged as being full sibs or avunculars if 0.17 ≤ Z0 ≤ 0.33, 0.33 < Z1 ≤ 0.6 and 0.17 ≤ Z2 ≤ 0.33. Half siblings or avunculars were defined as having 0.4 ≤ Z1 ≤ 0.6 and 0.33 < Z0 ≤ 0.6. Some of the duplicates flagged in this step were expected, being genotyped in multiple data sets and hence having the same cohort identifications (IDs). In this case, one of each pair was removed from the final data set. For instance, if two duplicates were flagged as unrelated, the closer relation was kept, whereas if two unrelated individuals were expected, the older of the two was kept to reduce the total number of samples.

After removing duplicate IDs and flagging related pairs of IDs, we used eigenstrat principal components, generating the following model:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \]

Where \( Y \) is an SNP genotypic value, \( \beta_0 \) is the intercept, \( \beta_1, \beta_2, \beta_3, \beta_4 \) are the regression coefficients for the principal components, and \( \epsilon \) is the error term. The resulting regression coefficients were used as covariates in the GWAS analysis to account for relatedness among samples.

Stage 2 phenotype categorization. Participants in both NHS and HPFS cohorts reported new SCC diagnosis biennially. With their permission, medical records were obtained and reviewed to confirm their self-reported diagnosis. Eligible cases in the NHS and HPFS consisted of participants with pathologically confirmed invasive SCC, diagnosed any time after baseline up to the 2012 follow-up cycle for both cohorts. Samples free of diagnosis of SCC were controls in this study. In the three compiled data sets, samples without information on SCC diagnosis were excluded. Among the samples with imputed genotypes, we have 367 SCC cases and 5,453 controls in Affymetrix data set, 220 SCC cases and 2,901 controls in Illumina HumanHap data set and 238 SCC cases and 3,164 controls in Illumina Omni Express data set—825 SCC cases and 11,518 controls in total.

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

\[ \text{SCC diagnosis} = \text{age} + \text{sex} + \text{BCC history} + p_1 + p_2 + p_3 + p_4 + p_5 + \text{genotype} \]

These principal components were calculated for all individuals on the basis of ~10,000 unlinked markers using the EIGENSTRAT software.

Proportion of familial relative risk. We have used the formula for calculating the proportion of familial relative risk (FRR) as outlined by the Cancer Oncological Environment Study (http://www.nature.com/icogs/primer/common-variation-and-heritability-estimates-for-breast-ovarian-and-prostate-cancers/#70) as described previously. The ORs derived from our meta-analysis of stage 1 and stage 2 are assumed to be relative risks. We estimated the proportion of the FRR explained by each SNP (FRR exp) as:

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

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

\[ \frac{1}{2} \sum \frac{1}{\lambda_i} \]

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:

\[ \log(\hat{\lambda}) / \log(\lambda_i) \]

In this equation, \( \hat{\lambda} \) is the familial relative risk observed in epidemiological studies.

Regulatory function of novel variants. For each novel SCC susceptibility variant, we searched for evidence of regulatory function using haploReg version 4 (refs 8, 44) (http://www.broadinstitute.org/tey/mammals/haploreg/haploreg.php).

P < 10^-8 in any of the quality control (QC) regressions and were removed. No SNPs in the Illumina Omni Express data set had P-values < 10^-6; 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 SNPs and the rest of the studies controls as ‘controls.’ For example, in the Illumina Omni Express data set, we ran regression analyses of NHS-gout controls considered as ‘cases’ versus the HPFS-gout, NHS-endometrial 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 < 10^-8) and examined them. If all possible SNPs were flagged as significant and 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
We queried each rsID and extracted data from ENCODE Project Consortium 2011–2012 on closest annotated gene, chromatin immunoprecipitation–sequencing (ChIP-seq) factor binding, DNaseI hypersensitivity sites, and enhancer and promoter chromatin segmentation states46–48. Data were also extracted from Roadmap Epigenomics Consortium 2015 on enhancer and promoter chromatin segmentation states, specifically using the following states: 15-state HMM, 25-state HMM, H3K4me3, H3K4me1, H3K27ac, and H3K9ac49. We particularly focused on the enhancer and promoter annotations that referenced normal human epidermal keratinocytes 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 many other studies50. We made special note of variants that were eQTLs in skin tissue.

Power calculations. Power was computed according to Freidlin et al.51. To account for misclassification, expected genotype frequencies in study cases were replaced with a mixture of genotype frequencies in true cases and in true controls. Power was plotted as a function of OR for detecting a variant with minor allele frequency 0.1, based on the GWAS sample size and with hypothetical misclassification rates of 0, 10 and 20% (where the specified fraction of study cases are misclassified controls).

Data availability. GWAS data from the 23andMe research cohort and Nurses’ Health Study have not been deposited in public repositories, as consent for this was not obtained in the study protocols. The precomputed rankings and P-values for SNPs included in the stage 1 GWAS are available upon request by contacting D.A.H. at dhinds@23andme.com. Precomputed rankings and P-values for the top 10,000 SNPs included in the stage 2 GWAS are freely available by contacting www.channing.harvard.edu/nhs. Any additional data (beyond those included in the main text and Supplementary Information) that support the findings of this study are available from the corresponding author upon request.

References
Acknowledgements

We thank the customers and employees of 23andMe for making this work possible. In addition, we thank Constance Chen for her technical support, and the participants and staff of the Nurses’ Health Study, the HPFS, 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, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA and WY. We 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 CA7969, UM1 CA186107, UM1 CA167552, Damon Runyan (J.Y.T.) the Stanford TRAM (Translational Research and Applied Medicine Program), and the Dermatology Foundation Career Development Award (K.Y.S.).

Author contributions

J.H., K.Y.S. and J.Y.T. designed and oversaw the study. Y.L., H.S.C., J.H., D.A.H. and K.Y.S. were responsible for quality control, manuscript writing and data analyses. J.H., P.K., W.-Q.L., A.A.Q. and D.A.H. contributed to data acquisition. K.J.R., J.Y.T., K.Y.S. and H.S.C. carried out survey validation collection and analysis. Bioinformatics analyses were carried out by H.S.C., Y.L., D.A.H., W.W., H.-J.D., J.H. and K.Y.S. Major contributions to writing and editing were especially made by H.S.C., Y.L., J.H., K.Y.S., D.A.H. and J.Y.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. All other authors declare no competing financial interests.

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

How to cite this article: Chahal, H. et al. Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. Nat. Commun. 7:12048 doi: 10.1038/ncomms12048 (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