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 States1. 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 RALY2,3. 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.

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 locus1. 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 × 10⁻³ and P = 1.3 × 10⁻², respectively, by logistic regression) (Table 3 and Supplementary Table 2).

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>Affy 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 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 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 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>479 (41.0)</td>
<td>5651 (49.1)</td>
<td>1148 (9.9)</td>
</tr>
<tr>
<td>Combined meta-analysis (n = 29,9840)</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.
Figure 1 | Manhattan plot of stage 1 GWAS analysis of SCC. Total stage 1 GWAS analysis included 6,579 cases and 280,558 controls from the 23andMe cohort. The y-axis represents log-scaled P-values. Loci with smallest \( P < 5 \times 10^{-8} \) (genome-wide significance) are shown in red. Seven SCC susceptibility loci reached genome-wide significance in stage 1, including six previously reported pigmentation loci at 6p25.3 (IRF4), 5p13.2 (SLC4A52), 16q24.3 (MCIR), 11q43.3 (TYR), 2q21.22 (RALY-ASIP) and 15q13.1 (OCA2), which are highlighted by asterisks. The seventh susceptibility locus reaching significance in stage 1, 9q34.3 (SEC16A), is novel and not pigmentation related (marked by a double asterisk). Three additional novel susceptibility loci, 2p22.3, 7p21.1 (name of the nearest gene. Positions with significance in stage 1, including six previously reported pigmentation loci at 6p25.3 (IRF4), 5p13.2 (SLC4A52), 16q24.3 (MCIR), 11q43.3 (TYR), 2q21.22 (RALY-ASIP) and 15q13.1 (OCA2), which are highlighted by asterisks. The seventh susceptibility locus reaching significance in stage 1, 9q34.3 (SEC16A), is novel and not pigmentation related (marked by a double asterisk). Three additional novel susceptibility loci, 2p22.3, 7p21.1 (AHR) and 11q23.3 (CADM1-BUD3), reached genome-wide significance in the overall meta-analysis, as did the previously reported locus 9p22.2 (BNC2-CNTLN) (Table 2). The remaining three loci in the figure, 3q28 (LPP), 6p21.3 (HLA-DQBI) and 8q23.3 (TRPS1), had \( P < 10^{-5} \) in stage 1 but did not reach genome-wide significance in the meta-analysis.

Table 2 | Eleven loci reaching genome-wide significance in two-stage GWAS of SCC.

<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>
</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>1.62</td>
<td>3.0 ( \times 10^{-10} )</td>
<td>1.60</td>
</tr>
<tr>
<td>rs1805007</td>
<td>16q24.3</td>
<td>MCIR</td>
<td>C/T</td>
<td>0.07 (1.0)</td>
<td>1.45</td>
<td>5.8 ( \times 10^{-43} )</td>
<td>1.64</td>
</tr>
<tr>
<td>rs35407</td>
<td>5p13.2</td>
<td>SLC4A52</td>
<td>G/A</td>
<td>0.04 (0.98)</td>
<td>0.59</td>
<td>3.6 ( \times 10^{-15} )</td>
<td>0.62</td>
</tr>
<tr>
<td>rs1126809</td>
<td>11q14.3</td>
<td>TYR</td>
<td>G/A</td>
<td>0.28 (0.99)</td>
<td>1.17</td>
<td>3.5 ( \times 10^{-14} )</td>
<td>1.08</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>1.28</td>
<td>3.6 ( \times 10^{-14} )</td>
<td>1.08</td>
</tr>
<tr>
<td>rs18090407</td>
<td>15q13.1</td>
<td>OCA2</td>
<td>C/T</td>
<td>0.07 (1.0)</td>
<td>1.21</td>
<td>8.0 ( \times 10^{-9} )</td>
<td>0.97</td>
</tr>
<tr>
<td>rs57994353*</td>
<td>9q34.3</td>
<td>SEC16A</td>
<td>T/C</td>
<td>0.30 (0.99)</td>
<td>1.12</td>
<td>1.4 ( \times 10^{-8} )</td>
<td>1.05</td>
</tr>
<tr>
<td>rs10810657</td>
<td>9p22.2</td>
<td>BNC2, CNTLN</td>
<td>A/T</td>
<td>0.41 (0.99)</td>
<td>0.91</td>
<td>2.4 ( \times 10^{-7} )</td>
<td>0.83</td>
</tr>
<tr>
<td>rs74899442*</td>
<td>11q23.3</td>
<td>CADM1, BUD3</td>
<td>T/C</td>
<td>0.01 (0.62)</td>
<td>2.12</td>
<td>3.0 ( \times 10^{-7} )</td>
<td>2.42</td>
</tr>
<tr>
<td>rs192481803*</td>
<td>2p22.3</td>
<td>unknown</td>
<td>C/T</td>
<td>0.01 (0.58)</td>
<td>1.92</td>
<td>4.5 ( \times 10^{-7} )</td>
<td>1.51</td>
</tr>
<tr>
<td>rs117132860*</td>
<td>7p21.1</td>
<td>AHR</td>
<td>G/A</td>
<td>0.02 (0.52)</td>
<td>1.46</td>
<td>7.2 ( \times 10^{-7} )</td>
<td>1.80</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. SNPs that met genome-wide significance (\( P < 5 \times 10^{-8} \)) via logistic regression in stage 1 and/or overall meta-analysis are listed. In addition, we report genetic locus, nearest gene, major allele, minor allele, MAF as calculated from stage 1 data, avg imp \( r^2 \) (a measure of imputation quality) for stage 1 and OR with \( P \)-value for each stage, calculated with respect to the minor allele. In stage 1, we analysed 6,579 SCC cases and 280,558 controls of European ancestry in the United States. In stage 2, we analysed 825 SCC cases and 11,518 controls of European ancestry in the United States. We then combined the data from these 2 stages (which resulted in 7,404 SCC cases and 292,076 controls) and performed fixed-effect meta-analysis. Statistics for effect heterogeneity (\( \theta_{cat} \) and \( I^2 \)) are included in Supplementary Tables 2 and 3. Asterisks highlight novel loci.

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, \(r^2 = 0.33, D' = 1\)), 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 pigmentation 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 \(\sim 13\) kb upstream of the basonuclin 2 (BN2C) transcription start site in a potential enhancer in foreskin melanocytes and other cell types\(^{14}\). This SNP is in linkage disequilibrium with rs12350739 (\(r^2 = 0.90, D' = 1.00\)), which resides in an enhancer element regulating BN2C transcription in human melanocytes\(^{14}\). BN2C is a DNA-binding zinc-finger protein thought to act as both a messenger RNA-processing enzyme and a transcription factor\(^{14}\). BN2C is expressed in melanocytes and, to a lesser extent, keratinocytes, with higher expression levels corresponding to darker skin pigmentation in human skin tissue analysis. Variants in the BN2C 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.

**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 SEC61A R1039C (\(r^2 = 0.90, D' = 1.00\)). SEC61A 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 SEC61A corresponding to increased secretion and cellular proliferation\(^{17}\). Overexpression of SEC61A 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 had significantly shorter median survival (36 versus 54 months)\(^{23}\). Conversely, overexpression of CADM1 in SCC cells suppresses cell proliferation and promotes apoptosis\(^{23}\). In light of these recent findings from various studies, our results provide further evidence that CADM1 may play a role in SCC development.

The third and fourth novel susceptibility variants rs192481803 and rs117132860 reached genome-wide significance in the overall meta-analysis. rs192481803 at 2p22.3 (\(P = 4.5 \times 10^{-8}\), OR = 1.90, logistic regression) lies within a long non-coding RNA AC012593.1 of unknown function\(^{23}\). rs117132860 (\(P = 3.6 \times 10^{-8}\), OR = 1.48, logistic regression), located at 7p21.1, is situated within a DNAse hypersensitivity site 203 kb upstream of AHR, the aryl hydrocarbon receptor, and has predicted enhancer activity in multiple tissues including keratinocytes. AHR is a widely expressed transcription factor that is involved in drug metabolism and cellular proliferation, although its exact role is tissue dependent and oftentimes paradoxical\(^{24-26}\). Interestingly, in the context of keratinocytes, ultraviolet radiation activates AHR and triggers a molecular cascade (known as the AHR-E2F1-CHK1 axis) that ultimately results in apoptosis\(^{24}\). This inhibition allows ultraviolet-induced DNA damage and reactive oxygen species to accumulate within keratinocytes; hence, overexpression of AHR may contribute to the development of non-melanoma skin cancers.

**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-DQ8.1. Transplant and other immunosuppressed patients are at significantly increased risk of SCCs suggesting a role for T-cell-mediated protection against SCC\(^{27}\). HLA-DQ8.1 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. rs1715549 is in tight linkage disequilibrium with SNPs associated with vitiligo (rs9851967, $r^2 = 0.93$), celiac disease (rs1465410, $r^2 = 0.74$) and allergy (rs9860547, $r^2 = 0.87$)29,30. rs199816436 lies inside an intron in TRPS1, a transcription factor that represses GATA-regulated genes and binds to a dynem light chain protein. rs199816436 is tightly correlated with rs10080475 ($r^2 = 0.85$) and is a TRPS1 expression quantitative trait locus (eQTL) in liver29. 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:

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

(1)

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 relatedness, using a graph-based method to correct for relatedness that persisted after principal component analysis; the genomic control inflation factor was 1.085 (computed from the median $P$-value for results that passed quality control). Regions of interest were defined 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 + BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The V3 platform was based on the Illumina OmniExpress + BeadChip, with custom content to improve the overlap with our V2 array, with a total of ~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 ~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 customer service to provide additional samples, as is done for all 23andMe customers. Individuals were only included if they had >97% European ancestry, as determined through an analysis of local ancestry31. Briefly, this analysis first partitions phased genomic data into short windows of ~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 classification is then used to infer the haplotype phase 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 datasets in the Human Comparison Diversity Project, HapMap and 1000 Genomes) and from 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 algorithm32. 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. Two individuals share either one or both genomic segments identical-by-descent if their IBD sharing exceeds the minimal expected sharing between first cousins in an outbred population.

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 survey responses. P-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 the 48 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. In the HPFS, we enrolled in 1986 when more than 50,000 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 mail. 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 genotype data available. We combined these data sets into three compiled data sets based on their genotype platform type: Affymetrix, Illumina HumanHap series or Illumina Omni Express. The Affymetrix data set comprised data on the Affy 6.0 and 6.1 arrays, while the Illumina data sets contained data on the Illumina HumanHap series, the Illumina HumanOmniExpress and the Illumina HumanHap data sets. In each data set, controls were genotyped in the 1986 questionnaire. For example, in the Illumina Omni Express data set, we ran regressions of NHS-gout status on their medical histories and baseline health-related exposures. 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 mail. 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 phenotype categorization.** Participants in both NHS and HPFS cohorts reported new SCC diagnosis biannually. 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 ProBABEL 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. The 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.

For Z ≤ 0, Z ≤ 0.1 and 0.1 < Z ≤ 0.9, we flagged as being identical twins or duplicated IDs. We considered full siblings if 0.17 ≤ Z ≤ 0.33, 0.34 ≤ Z ≤ 0.6 and 0.6 < Z ≤ 0.9. Half siblings or avunculars were defined as having 0.4 ≤ Z ≤ 0.6 and 0.7 < Z ≤ 0.9. Some of the duplicates flagged in this step were expected, having been genotyped in multiple data sets and hence having the same cohort identifications (IDs). In this case, one of each pair was chosen for removal from secondary genotype data analysis. In other instances where genotypes were flagged as unexpected duplicates with the different cohort IDs, but pairwise genotype concordance rate > 0.999, resulted in removal of both individuals from the pair. Related individuals (full sibs and 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 8,065 individuals. Of the 6,894 individuals originally in the Illumina data set 107 were removed, because they were duplicates or flagged for removal in the genotyping step, leaving 7,687 IDs. In addition, eight pairs of individuals were flagged as related. In the Omni express data set, there were 3,036 individuals at the start with 39 IDs removed, leaving 3,917 IDs and 5 pairs of related IDs. After removing duplicate IDs and flagging related pairs of IDs, we used eigenstrat to run principal component analysis on each compiled data set, removing one member from each flagged pair of related individuals. For Affymetrix and Illumina HumanHap, we used ~12,000 SNPs that were filtered, to ensure low pairwise linkage disequilibrium (LD)28. For the OmniExpress data set, we used ~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 status on their medical histories and baseline health-related exposures. Between 1993 and 1994, 18,159 study participants provided blood samples by overnight mail. Information on SCC development was first collected in the 1986 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 mail. 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.

We then ran a pairwise IBD analysis for each combined data set to detect duplicate and related individuals based on resulting Z-scores. For Z ≤ 0, Z ≤ 0.1 and 0.1 < Z ≤ 0.9, we flagged as being identical twins or duplicated IDs. We considered full siblings if 0.17 ≤ Z ≤ 0.33, 0.34 ≤ Z ≤ 0.6 and 0.6 < Z ≤ 0.9. Half siblings or avunculars were defined as having 0.4 ≤ Z ≤ 0.6 and 0.7 < Z ≤ 0.9. Some of the duplicates flagged in this step were expected, having been genotyped in multiple data sets and hence having the same cohort identifications (IDs). In this case, one of each pair was chosen for removal from secondary genotype data analysis. In other instances where genotypes were flagged as unexpected duplicates with the different cohort IDs, but pairwise genotype concordance rate > 0.999, resulted in removal of both individuals from the pair. Related individuals (full sibs and 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 8,065 individuals. Of the 6,894 individuals originally in the Illumina data set 107 were removed, because they were duplicates or flagged for removal in the genotyping step, leaving 7,687 IDs. In addition, eight pairs of individuals were flagged as related. In the Omni express data set, there were 3,036 individuals at the start with 39 IDs removed, leaving 3,917 IDs and 5 pairs of related IDs. After removing duplicate IDs and flagging related pairs of IDs, we used eigenstrat to run principal component analysis on each compiled data set, removing one member from each flagged pair of related individuals. For Affymetrix and Illumina HumanHap, we used ~12,000 SNPs that were filtered, to ensure low pairwise linkage disequilibrium (LD)28. For the OmniExpress data set, we used ~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 status on their medical histories and baseline health-related exposures. Between 1993 and 1994, 18,159 study participants provided blood samples by overnight mail. Information on SCC development was first collected in the 1986 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 mail. 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.

**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 Gene-environment Study (http://www.nature.com/icogs/primer/common-variation-and-heritability-estimates-for-bowel-ovarian-and-prostate-cancers/#70) as described previously25. 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\_SNP) as:

\[
\text{FRR}_{SNP} = \frac{(p^2 - q^2)}{(p^2 + q^2)}
\]

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 population database. Assuming that the loci combine multiplicatively and are not in linkage disequilibrium, the combined effect of all loci is given by:

\[
\frac{1}{Z} = \sum_{i=1}^{n} \lambda_i
\]

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:

\[
\log \left( \frac{\lambda_{Z}}{\lambda_{0}} \right)
\]

In this equation, \(\lambda_{0}\) is the familial relative risk observed in epidemiological studies. \(\lambda_{Z}\) is fourfold for SCC43.

**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/annotation/haploreg/haploreg.php)45.


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 CA87969, U1U1 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