Quantifying the Polygenic Contribution to Cutaneous Squamous Cell Carcinoma Risk

Joanne E. Sordillo¹, Peter Kraft², Ann Chen Wu¹,4 and Maryam M. Asgari¹,3,4

Genetic factors play an important role in cutaneous squamous cell carcinoma risk. Genome-wide association studies have identified 21 single nucleotide polymorphisms associated with cutaneous squamous cell carcinoma risk. Yet no studies have attempted to quantify the contribution of heritability to cutaneous squamous cell carcinoma risk by calculating the population attributable risk using a combination of all discovered genetic variants. Using an additive multi-locus linear logistic model, we determined the cumulative association of these 21 genetic regions to cutaneous squamous cell carcinoma population attributable risk. We computed a multi-locus population attributable risk of 62%, suggesting that if the effects of all the risk alleles were removed from a population, the cutaneous squamous cell carcinoma risk would drop by 62%. Using stratified analysis, we also examined the impact of sex on polygenic risk score, and found that men have an increased relative risk throughout the spectrum of the polygenic risk score. Quantifying the impact of genetic predisposition on the proportion of cancer cases can guide future research decisions and public health policy planning.

INTRODUCTION

Cutaneous squamous cell carcinoma (cSCC) is a common form of skin cancer responsible for a substantial public health burden and significant health care costs (Housman et al., 2003). Non-Hispanic whites have a high lifetime prevalence of cSCC, ranging from 7% to 11% (Kallini et al., 2015). Previous studies have shown that cSCC risk clusters in families, suggesting a strong heritable component to this common form of skin cancer (Asgari et al., 2015; Hussain, 2009), but the contribution of genetics to cSCC risk has not been quantified. Recently, genome-wide association studies (GWAS) among non-Hispanic whites have identified variants in 21 genetic loci that are associated with cSCC risk (Asgari et al., 2016; Chahal et al., 2016; Siiskonen et al., 2016). But the degree to which these genetic variants impact the overall burden of cSCC is unclear.

One way to measure the contribution of genetic variants to disease risk is to examine the population attributable risk (PAR), which approximates the reduction in incidence that would be achieved if the risk factor was eliminated from that population (Rothman, 2008). Although risk variants cannot be removed from the population, genetic PARs are often used in epidemiologic research to estimate the degree to which a disease can be attributed to the risk variant. The validity of the PAR for measuring genetic contributions to disease risk has been established (Witte et al., 2014) and has bounds that range from 0% to 100%. We computed a multi-locus PAR for cSCC using 21 published single nucleotide polymorphisms (SNPs) from large-scale GWAS. To examine how PAR can be modified by a patient variable, we explored the effects of sex on PAR by stratifying our population by sex. Finally, we used data from the 21 published variants to calculate a polygenic risk score. In other cancers, such as breast cancer, risk prediction models that combine a polygenic risk score with epidemiologic risk factors provide substantial risk stratification of the general population (Maas et al., 2016; Meads et al., 2012). Quantifying the genetic contribution of these SNPs to cSCC risk, in combination with clinical and environmental risk factors, could help identify individuals at greatest risk of developing cSCC, who would benefit most from enhanced monitoring programs.

RESULTS

Cutaneous SCC risk was associated with 21 SNPs, all of which were used to compute a multi-locus PAR (Kraft et al., 2009; Rockhill et al., 1998). Not all risk alleles are minor alleles. For SNPs associated with a reduction of cSCC risk (“protective” SNP), we took the inverse odds ratio and the major allele frequencies for the association. This way, all variants could be added to the population attributable risk calculation. The multi-locus PAR (Table 1) suggests that if the effects of all risk alleles were removed from a population, there would be a 62% reduction in cSCC cases. SNPs in IRF4, TYR, RALY, FOXP1, and MCIR used to compute the multi-locus PAR were identified in more than one GWAS, and had similar odds ratios, showing a consistency across study populations. Prevalent SNPs with moderate to high odds ratios contributed the most the multi-locus PAR, and
included SNPs in the genes \textit{SLC45A2}, \textit{SRC}, \textit{HERC2}, \textit{DEF8}, and \textit{HLADQA1}.

Figure 1 shows the relative risk of cSCC increasing with higher percentiles of the polygenic risk score, based on 21 risk alleles. Sex-specific curves for males and females are shown, which reveal that males have higher risk as compared to females, particularly at polygenic risk scores that exceed the 40th percentile. This suggests a potential interaction between sex and genetic score. The risk of cSCC is twice the average population risk at the 84th percentile of the polygenic risk score for females and at the 77th percentile for males. A two-fold increased relative risk is a threshold widely used as a benchmark for clinically meaningful increased risk for common diseases, such as cSCC (Roberts et al., 2012).

\textbf{DISCUSSION}

Different measures can be used to assess how much known genetic factors contribute to overall cSCC risk including heritability on various scales, sibling relative risk, log relative risk genetic variance, the area under the receiver operating curve, and the PAR (Witte et al., 2014). The utility and limitations of these various metrics have been discussed extensively elsewhere (Witte et al., 2014). Here, we focused on the PAR and log relative risk variance (which is used to calculate the population distribution of the polygenic risk score) because of their direct relevance to the clinical utility of these genetic variants for cSCC prevention and screening.

Most polymorphisms that contribute to the multi-locus PAR are within biologically plausible gene candidates for cSCC etiology. \textit{IRF4} (Han et al., 2008) is associated with pigmentation phenotypes. Variants in \textit{DEF8} are associated with expression of a cell cycle progression gene (\textit{CDK10}) in sun-exposed skin (GTEx Consortium, 2015). Expression of \textit{CADM1}, a gene that modifies tumor interaction with cell-mediated immunity, is associated with survival in cSCC patients (Liu et al., 2013). Activation of \textit{AHR} during UV radiation exposure may decrease apoptosis in keratinocytes (Frauenstein et al., 2013) with potential consequences for enhancing cancer risk. \textit{ERBB2IP} has been shown to affect the Ras pathway (Kolch, 2003). Risk alleles in \textit{SLC45A2}, a gene associated with skin pigmentation, were also linked to cSCC risk.

While the multi-locus PAR allows quantification of the joint impact of these variants on cSCC risk, this one statistic is of limited utility for screening purposes, as it does not consider known environmental risk factors, such as sun exposure and smoking. However, calculation of a polygenic risk score, using the combination of these risk alleles, helps us to visualize their relationship with cSCC risk, and may be a useful metric, when combined with relevant clinical variables, for screening. Although we have demonstrated the potential utility of the polygenic risk score here, this type of screening is not yet at a stage where it could be incorporated into clinical practice. The next phase of risk prediction models should incorporate both clinical and environmental risk factors, combined with a polygenic risk score to help optimize primary and secondary prevention strategies for cSCCs. Ideally, the model would be developed and calibrated in one population, and validated in an independent population. Identification of individuals with substantially elevated risk is of paramount importance for early detection and prevention efforts.

Table 1. Cutaneous squamous cell carcinoma genome-wide association studies variants contribute to a multi-locus population attributable risk of 62%

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>SNP</th>
<th>Risk Allele</th>
<th>Risk Allele Frequency</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>2</td>
<td>rs192481803</td>
<td>T</td>
<td>0.01</td>
<td>1.9</td>
</tr>
<tr>
<td>FOXP1</td>
<td>3</td>
<td>rs62246017</td>
<td>A</td>
<td>0.33</td>
<td>1.06</td>
</tr>
<tr>
<td>TPRG1/TIF63</td>
<td>3</td>
<td>rs6791479</td>
<td>T</td>
<td>0.43</td>
<td>1.05</td>
</tr>
<tr>
<td>ERBB2IP</td>
<td>5</td>
<td>rs17247181</td>
<td>T</td>
<td>0.10</td>
<td>1.40</td>
</tr>
<tr>
<td>SLC45A2</td>
<td>5</td>
<td>rs35407</td>
<td>G</td>
<td>0.96</td>
<td>1.69</td>
</tr>
<tr>
<td>IRF4</td>
<td>6</td>
<td>rs12203592</td>
<td>T</td>
<td>0.17</td>
<td>1.62</td>
</tr>
<tr>
<td>HLADQA1</td>
<td>6</td>
<td>rs4455710</td>
<td>T</td>
<td>0.38</td>
<td>1.17</td>
</tr>
<tr>
<td>PARK2</td>
<td>6</td>
<td>rs9689649</td>
<td>C</td>
<td>0.22</td>
<td>1.38</td>
</tr>
<tr>
<td>AHR</td>
<td>7</td>
<td>rs17132860</td>
<td>A</td>
<td>0.02</td>
<td>1.48</td>
</tr>
<tr>
<td>ST3GAL1</td>
<td>8</td>
<td>rs9643297</td>
<td>G</td>
<td>0.31</td>
<td>1.21</td>
</tr>
<tr>
<td>SEC16A</td>
<td>9</td>
<td>rs57994353</td>
<td>C</td>
<td>0.30</td>
<td>1.12</td>
</tr>
<tr>
<td>BNC2, CNTLN</td>
<td>9</td>
<td>rs10810657</td>
<td>A</td>
<td>0.59</td>
<td>1.11</td>
</tr>
<tr>
<td>TYR</td>
<td>11</td>
<td>rs1126809</td>
<td>A</td>
<td>0.28</td>
<td>1.16</td>
</tr>
<tr>
<td>CADM1-BUD13</td>
<td>11</td>
<td>rs74899442</td>
<td>C</td>
<td>0.01</td>
<td>2.13</td>
</tr>
<tr>
<td>OCA2</td>
<td>15</td>
<td>rs1800407</td>
<td>T</td>
<td>0.07</td>
<td>1.2</td>
</tr>
<tr>
<td>HERC2</td>
<td>15</td>
<td>rs12916300</td>
<td>T</td>
<td>0.74</td>
<td>1.14</td>
</tr>
<tr>
<td>DEF8</td>
<td>16</td>
<td>rs4268748</td>
<td>C</td>
<td>0.26</td>
<td>1.33</td>
</tr>
<tr>
<td>MCHR</td>
<td>16</td>
<td>rs1805007</td>
<td>T</td>
<td>0.07</td>
<td>1.46</td>
</tr>
<tr>
<td>DEF8</td>
<td>16</td>
<td>rs8063761</td>
<td>T</td>
<td>0.33</td>
<td>1.34</td>
</tr>
<tr>
<td>RALY-ASIP</td>
<td>20</td>
<td>rs6059655</td>
<td>A</td>
<td>0.07</td>
<td>1.27</td>
</tr>
<tr>
<td>SRC</td>
<td>20</td>
<td>rs754626</td>
<td>G</td>
<td>0.25</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Abbreviation: SNP, single nucleotide polymorphism.

1These SNPs have been reported as protective SNPs, thus, the major risk allele and its corresponding odds ratio (the reciprocal of the protective odds ratio) are shown.
The PAR is commonly used to approximate the public health implications of modifying or removing an exposure. Although it is a useful quantifiable estimate of the impact of a causal factor (in this case, genetic risk), it can rapidly approach the upper bound of 100%, if the risk allele frequency and relative risk of the disease are high, and may be more inflated than other measures of genetic risk, such as the heritability of disease liability, approximate heritability, sibling recurrence risk, and overall genetic variance using a log relative risk scale (Witte et al., 2014). The PAR is based on the excess fraction of disease that is associated with the presence of these risk alleles, but not necessarily the etiologic fraction, or the fraction of disease that truly arises due to these risk alleles. An additional limitation of the multi-locus PAR and the polygenic risk score based on these loci is that they do not account for interactions among SNPs. Furthermore, it should be noted that these GWAS were done primarily in non-Hispanic white populations so these risks may not translate to other ethnic groups.

Despite these limitations, the multi-locus population attributable risk and its related polygenic risk score are promising tools for identifying those individuals at greatest risk for cSCC, who would benefit from enhanced monitoring, and those individuals who are at lowest cSCC risk, for whom enhanced monitoring would be time-consuming, costly, and unnecessary. Development of such a risk prediction tool for cSCC would benefit from further refinement that includes gene–environment interactions and from validation across different populations, but would have the potential to be clinically meaningful, impacting both screening and prevention efforts. Next-generation studies relating the polygenic contribution to cSCC risk will likely incorporate newly discovered rare variants. Future studies may also include variants identified through meta-analyses of existing as well as forthcoming cSCC GWAS, which will increase both the power and robustness of genetic associations used to derive the polygenic risk score.

**MATERIALS AND METHODS**

We performed a search of the published literature using standard search strategies involving the querying of two online databases (MEDLINE and Cochrane) using key words squamous cell carcinoma, skin, and genome-wide association from January 1980 to January 31, 2017, followed by evaluation of the bibliographies of relevant articles, and identified three GWAS that utilized cohorts from Kaiser Permanente Northern California, Nurses’ Health Study, the Health Professionals Follow-up Study, the Rotterdam Study, and the 23andMe (Asgari et al., 2016; Chahal et al., 2016; Siiskonen et al., 2016). We excluded one study that was primarily a basal cell carcinoma GWAS that tested significant basal cell carcinoma–associated SNP variants for their relationship with cSCC, and was not a true SCC GWAS (Nan et al., 2011). Whereas our methodology was not a formalized systematic review, we did capture all published cSCC GWAS data. We compiled all SNPs associated with cSCC that replicated in independent populations. Odds ratios for SNPs associated with cSCC were used for the multi-locus PAR calculations, and all studies assumed an additive genetic model. We focused on bi-allelic SNPs, removing one tri-allelic SNP, duplicates, or SNPs in linkage disequilibrium. If the same SNP was reported in more than one GWAS, or if two SNPs at the same locus were in linkage disequilibrium ($R^2 > 0.3$), we used the odds ratio from the largest study for the multi-locus PAR and polygenic risk score calculations. Odds ratios for individual risk alleles, as well as the multi-locus PAR for all alleles combined, are shown in Table 1. The multi-locus PAR is weighted most heavily by the risk alleles with both the highest prevalence and largest odds ratios for cSCC. We used the formula below (derived from Rockhill et al., 1998) to compute the multi-locus PAR:

$$1 - \frac{1}{1 + \prod_i \left( \sum_j p_i R_{i|j} \right)}$$

where $i$ indexes SNP, and $j$ indexes genotype at each SNP.

We also estimated the population distribution of polygenic risk scores based on published cSCC GWAS. For risk prediction based on multiple loci, we assumed a log-additive model for the joint effects of SNPs and constructed polygenic risk scores by summing the number of alleles across SNPs:

$$\log(\text{relative risk}) = \alpha + \sum_i \beta_i G_i$$

Here $\beta_i$ is the log odds ratio per risk allele at SNP $i$ (given in Table 1) and $G_i$ is the count of risk alleles at locus $i$ and $\alpha$ is chosen so that the average relative risk is 1.

The standard deviation of the polygenic risk score was calculated as follows,

$$\sqrt{\sum_i 2q_i (1 - q_i) \beta_i^2}$$

where $\beta_i$ is log relative risk ~ log odds ratio for SNP, and $q_i$ is (1 - risk allele frequency). For SNPs that were reported as protective, we used the major allele as the risk allele in our calculations, along with the corresponding cSCC odds ratio for the major allele. The standard deviation of the polygenic risk score can be used to plot the distribution of the relative risk due to known common SNPs: Relative risk of cSCC per polygenic risk score percentile is shown in Figure 1, with lines plotted for both males and females. We have plotted percentiles of the polygenic risk score in relation to cSCC risk, rather than the raw values themselves, to facilitate interpretation of the score relative to the population. The plot for males accounts for the underlying increased risk of cSCC in male subjects (relative risk = 1.31 for males vs. females) (Whiteman et al., 2016).

**CONFLICT OF INTEREST**
The authors state no conflicts of interest.
ACKNOWLEDGMENTS
This work was supported by the National Institutes of Health (R01 CA166672). Maryam Asgari has received unrelated grant funding to Massachusetts General Hospital from Valeant Pharmaceuticals, but the topic is not relevant to this work.

REFERENCES