Gene-environment analyses in a UK Biobank skin cancer cohort identifies important SNPs in DNA repair genes that may help prognosticate disease risk

Richie Jeremian[1,2], Pingxing Xie[1,2], Misha Fotovati[1,3], Philippe Lefrançois[1,3], Ivan V. Litvinov[1,2]

[1]Faculty of Medicine and Health Sciences, McGill University

[2]Department of Medicine, Division of Dermatology, Research Institute of the McGill University Health Centre (RI-MUHC) Montreal, Quebec

[3]Department of Medicine, Division of Dermatology, Lady Davis Institute (LDI), Jewish General Hospital, Montreal, Quebec

Abstract

Background: Despite well-established relationships between sun exposure and skin cancer pathogenesis/progression, specific gene-environment interactions in at-risk individuals remain poorly-understood.

Methods: We leveraged a UK Biobank cohort of basal cell carcinoma [BCC, n=17,221], cutaneous squamous cell carcinoma [cSCC, n=2,331]), melanoma *in-situ* (M-is, n=1,158), invasive melanoma (M-inv, n=3,798) and healthy controls (n=448,164) to quantify the synergistic involvement of genetic and environmental factors influencing disease risk. We surveyed 8,798 single nucleotide polymorphisms (SNPs) from 190 DNA repair genes, and 11 demographic/behavioral risk factors.

Results: Clinical analysis identified darker skin (RR=0.01~0.65) and hair (RR=0.27~0.63) colors as protective factors. Eleven SNPs were significantly associated with BCC, three of which were also associated with M-inv. Gene-environment analysis yielded 201 SNP-environment interactions across 90 genes (FDR-adjusted $q<0.05$). SNPs from the *FANCA* gene showed interactions with at least one clinical factor in all cancer groups, of which three (rs9926296, rs3743860, rs2376883) showed interaction with nearly every factor in BCC and M-inv.

Conclusions: We identified novel risk factors for keratinocyte carcinomas and melanoma, highlighted the prognostic value of several *FANCA* alleles among individuals with a history of sunlamp use and childhood sunburns, and demonstrated the importance of combining genetic and clinical data in disease risk stratification.

Impact: This study revealed genome-wide associations with important implications for understanding skin cancer risk in the context of the rapidly-evolving field of precision medicine.

Corresponding Author: Ivan V. Litvinov, McGill University Health Centre (MUHC), 1001 Decarie Boulevard, Montréal, QC, H4A 3J1, (514) 934-1934 ext. 76140, ivan.litvinov@mcgill.ca.

Conflicts of Interest: The authors declare no potential conflicts of interest.

Major individual factors (including sex, hair and skin color, and sun protection use) were significant mediators for all skin cancers, interacting with >200 SNPs across four skin cancer types.

Keywords

Introduction

Melanoma and keratinocyte carcinomas (KCs) (basal cell carcinoma (BCC) and cutaneous squamous cell carcinoma (cSCC)) constitute the majority of skin cancers, and collectively affect approximately >3 million individuals per year in the United States alone[1]. The incidence of these malignancies has been increasing over the past decades[2, 3].

Skin cancers arise from the interactions of numerous risk factors, including lifetime UV exposure, history of sun burns, fair skin, advancing age, and genetic predisposition[4–7]. Despite the well-established relationship between sun exposure and skin cancer pathogenesis, the interactions between specific genetic variants and sun exposure in predisposed individuals remain largely unknown. A particular genetically-determined pathway of interest involves UV-induced DNA damage, which promotes cells to enter a hyperproliferative (precancerous) state as a result of aberrant DNA repair mechanisms[8–12]. The impact of this interaction on the development of skin cancer is thought to be mediated by inherited genetic factors that impact the efficiency of such mechanisms. Taken together, these findings necessitate precise characterization of the synergistic contributions of both DNA repair-associated changes and environmental/behavioral factors on the predisposition toward these cancers.

Recent studies have demonstrated the importance of combining genetic and clinical data, often from large biomedical databases, to predict disease risk and facilitate drug design[13–15]. Notably, 23andMe®, AncestryDNA® and other direct-to-consumer platforms have already undertaken limited risk assessments for diseases on participants' genetic and phenotypic information and licensed these data for the development of a drug used to treat inflammatory skin conditions[16]. According to 23andMe® reports, more than 12 million kits were sold to consumers, while AncestryDNA® sold over 10 million kits. These commercialized genome-wide association study (GWAS)-based analyses sold directly to consumers have provided prediction of health risk for diabetes, Parkinson's, celiac disease, and many other diseases and conditions. Given the demonstrated limitations of using genetic data in isolation and the projected exponential increase in availability of diverse patient data, it is important to explore whether SNP analysis alone is sufficient for quantifying disease risk, and whether such analysis should be combined with demographic, behavioral and clinical information especially in dermatology[17, 18]. As this technology is used by millions of patients, oncologists/dermatologists need to understand and lead the effort of using GWAS appropriately to identify valuable predictive genetic variants in the context of clinical information.

To investigate these relationships in greater depth, we leveraged a dataset of four skin cancers and matched healthy controls from the UK Biobank (UKBB) database and performed associations between disease status, demographic and behavioral factors related to the impact of UV exposure, and single nucleotide polymorphism (SNP) genotypes across selected 190 DNA repair genes.

Materials and Methods

The study design and reporting followed the *STrengthening the REporting of Genetic Association Studies (STREGA): An Extension of the STROBE* Statement[19].

Participant demographics and genetic data

The UKBB is a large-scale biomedical database that comprises approximately 500,000 adult participants, aged 40 to 69, recruited between 2006 and 2010, and contains detailed genetic and health information collected during baseline and follow-up visits[20]. We included participants who self-reported their skin color during the initial UKBB assessment questionnaire. UKBB cancer registry data was obtained through linkage to UK national cancer registries and coded using the 10th revision of the International Classification of Diseases [ICD-10]. We first identified the participants with a diagnosis of either "malignant melanoma of skin" (C43 of ICD-10), or "other malignant neoplasms of skin" (C44 of ICD-10). To confirm a diagnosis of skin cancer, we required that the cancer had a histological type of BCC, cSCC, or melanoma *in-situ* (M-is) *vs.* melanoma invasive (M-inv).

For participants meeting the above criteria, as well as healthy subjects matched for self-reported skin tone, we extracted data from the "sun exposure" category, which includes the following items: "use of sun/UV protection (do not go out in sunshine, never/rarely, sometimes, most of the time, and always)", "number of childhood sunburns", "frequency of sunlamp use (number per year)", "time spent outdoors in summer (hours per day)" and "time spent outdoors in winter (hours per day)". We also extracted information regarding sex, age, "skin color (very fair, fair, light olive, dark olive, brown, black)", "hair color (red, blonde, light brown, dark brown, black)", "subjective age appearance (younger than expected for age, as expected for age, older than expected for age)", and "ease of skin tanning (never tan/only burn, mildly or occasionally tanned, moderately tanned, get very tanned)".

We also obtained single nucleotide polymorphism (SNP) data generated by the Affymetrix UK Biobank Axiom microarray. Based on a recent study that identified numerous genes directly and indirectly involved in various DNA repair pathways[21], we extracted SNP data (8,798 total markers) corresponding to 190 of these genes (Table S1), in addition to data on genetic principal components, genetic ethnic grouping, and genetic kinship to other participants, filtering exclusively for individuals identified demographically as Caucasian.

Quality control of genetic data

Extracted SNP data underwent quality control using PLINK version 2.00 (http://pngu.mgh.harvard.edu/purcell/plink/)[22]. We removed SNPs missing in more than 1% of individuals, SNPs with a minor allele frequency (MAF) less than 1%, SNPs falling outside Hardy-Weinberg Equilibrium by a threshold of $p = 0.000001$, and SNPs that lie in areas of

high linkage disequilibrium. We also removed individuals with a less than 99% SNP call rate, duplicate individuals and first-degree relatives. Following quality control, 1,912 SNPs remained for downstream analysis.

Following these selection criteria, we attained a cohort consisting of a total 472,672 individuals clustering into the following five groups: BCC (n=17,221), cSCC (n=2,331), melanoma *in-situ* (M-is, n=1,158), invasive melanoma (M-inv, n=3,798) and healthy control individuals (n=448,164). Within the skin cancer cohort, 1,288 individuals were identified as having greater than one disease diagnosis. Downstream analyses were conducted in a groupwise (disease *versus* control) fashion, independently for each disease group.

Disease-environment association analysis

Using the RStudio (version 2022.02.1, Build 461; https://www.rstudio.com/; RRID:SCR_000432) package *nnet*, we performed multinomial logistic regression to explore the relationship between the above-mentioned demographic and environmental factors, and risk of BCC, cSCC M-is and M-inv(23). These factors were set as predictor variables of outcome (disease), using one category in each variable as reference (sex – female; skin color – very fair; hair color – red; ease of skin tanning – never tan/only burn; subjective age appearance – younger than expected for age; use sun/UV protection – never/rarely). Factors composed of continuous variables (age, frequency of sunlamp use, number of childhood sunburns and time spent outdoors (summer and winter) were not transformed. The *relevel* function was used to select the healthy control group as the baseline outcome group, and analysis was performed using the *multinom* function for each predictor variable. The log(odds ratios) of each analysis were exponentiated and transformed into relative risk (RR) ratios, defined as the probability of disease for a given predictor variable compared to its reference (RR > 1 indicates increased disease risk; RR < 1 indicates decreased disease risk; RR ~ 1 indicates no difference in risk between disease and healthy control). Significance of risk was determined using a two-tailed Wald Z test(24).

SNP-disease and SNP-environment analyses

In order to assess how sun exposure factors interact with genetic susceptibility loci associated with skin cancers, we also performed gene-by-environment analysis using the *Robust-joint-interaction* package (https://epstein-software.github.io/robust-joint-interaction/), which computes *p*-values corresponding to tests of SNP and SNP-environment interactions(25). Multiple testing correction was performed using the False Discovery Rate approach ($q < 0.05$)(26). Significant interactions for each group were visualized as waterfall plots using the *GenVisR* package (https://bioconductor.org/packages/release/bioc/html/GenVisR.html)(27). Overlaps between significant SNP-environment interacting markers between each disease group were plotted using The *Molbiotools Multiple List Comparator* (https://molbiotools.com/listcompare.php) was used to plot overlaps between significant SNP-environment interactions (Venn Diagram) and assess for pairwise interactions between individuals with greater than one cancer diagnosis (Jaccard Index Coefficient).

Results

Disease-environment/behavior interactions

Demographic characteristics of our participant cohorts are summarized in Table 1. Using multinomial logistic regression(23), we computed ratios corresponding to skin cancer risk for each of our chosen demographic and environmental variables, using our four disease groups as outcome variables. We observed significant ($p<0.05$) trends corresponding to six demographic factors and environmental exposures (sex, hair color, skin color, ease of skin tanning, subjective age appearance, use of sun protection) for each disease group. BCC was positively correlated with male sex (RR=1.23) and frequent use of sunscreen (RR=2.40); and inversely correlated with light olive, dark olive or brown skin (RR=0.47~0.50), ease of tanning (RR=0.66), dark brown hair (RR=0.5), and older than expected appearance (RR = 0.79). cSCC was positively correlated with male sex (RR=1.97), older-than-expected appearance (RR=1.28), frequent ("always") use of sunscreen (RR=2.26); and inversely correlated with light olive, dark olive or brown skin (RR<0.33), ease of tanning (RR=0.58), as well as light brown, blonde, dark brown hair and black hair (RR=0.38~0.68). M-is was directly correlated occasional, frequent and very frequent use of sunscreen (RR=1.87~3.58); and inversely correlated with light olive, dark olive and brown skin (RR=0.46~0.65), blonde, light brown, dark brown and black hair (RR<0.60), and ease of tanning (RR=0.59). M-inv was positively correlated with occasional, frequent and very frequent use of sunscreen (RR=1.36~3.92); and inversely correlated with fair, light olive, dark olive and brown skin (RR=0.17~0.58), as well as blonde, light brown, dark brown and black hair (RR=0.27~0.72). These findings are summarized graphically in Figure 1 and numerically in Table S2.

SNP-disease association

We conducted SNP-disease association analysis across four UKBB skin cancer populations and identified eleven significant ($q<0.05$) alleles both positively and negatively associated with risk for BCC and M-inv (odds ratio [OR] 0.81–1.20). Through this analysis, we do not report any significant markers associated with cSCC and M-is. Out of 190 surveyed genes, only seven (Fanconi anemia complementation group A [*FANCA*], breast cancer 2 [*BRCA2*], RAD51 recombinase paralog B [*RAD51B*], Ribonucleotide reductase regulatory TP53 inducible subunit M2B [*RRM2B*], General transcription factor IIH subunit 4 [*GTF2H4*], exonuclease 1 [*EXO1*], and DNA cross-link repair 1B [*DCLRE1B*]) were highlighted in our analysis as being significantly associated with these cancers (Table 2). Three of the SNP markers (rs9926296, rs3743860, rs2376883) were found in BCC and M-inv groups and map to the intron regions *FANCA*, and are among the top positively (rs3743860, OR=1.20) and negatively associated SNPs with M-inv (rs9926296, OR=0.81). Moreover, two markers were in exon regions and are among the top three with positive association with BCC (rs4149909, *EXO1,* OR=1.13; rs61748588, *GTF2H4,* OR=1.19).

SNP-environment interaction

Since SNP analysis alone showed paucity of associations, we performed a robust joint test to assess the interaction between patient demographic, environmental/behavioral patterns and our surveyed DNA repair-associated SNPs. By adding this critical information, we identified 201 significant ($q<0.05$) SNPs localized in 104 genes that interact with demographic and

environmental/behavioral factors across four skin cancer groups. In the BCC group, we identified 32 such SNPs across 20 genes, 28 of which are in intron regions and 4 in exon regions. Notably, 17 of these 32 SNPs interacted with sunlamp use and 16 interacted with history of childhood sunburns. Hours spent outdoors, age, hair color and use of sun protection were also important categories, with 9 SNP-variable interactions in each group. Notably, markers rs12046289 (*DCLRE1B*), rs9926296 (*FANCA*) and rs3743860 (*FANCA*) interacted with every demographic and environmental/behavioral variable, while rs4942486 (*BRCA2*), rs2376883 (*FANCA*) and rs62989960 (*FANCA*) interacted with all but one variable.

In the cSCC group, we identified 72 significantly interacting SNPs across 54 genes, 59 of which are in introns and 13 in exons. By contrast to BCC, most of these SNPs (67 markers) were associated with sunlamp use, with 64 markers interacting exclusively with this variable. Moreover, rs17882704 (*CHEK2*) was the only SNP that interacted with greater than two variables, including sunlamp use, sex, tanning behavior, history of childhood sunburns and aging appearance. In the cSCC group, we did not identify any SNPs that interacted significantly with time spent outdoors, age, hair color and use of sun protection. One SNP (rs79594681 [*RAD51B*]) was conserved between both keratinocyte cancer groups and interacted with sunlamp use in both groups. Similarly, two genes (*MNAT* and *RECQL*) represented by distinct SNPs, were conserved across both BCC and cSCC. In the M-is group, we identified 106 significantly interacting SNPs across 69 genes, 90 of which were found in introns with the remaining 16 located in exon regions. Like cSCC, sunlamp use was the variable underpinning the majority (88 markers) of SNP interactions. History of childhood sunburns was the second most frequent variable, representing 16 SNP-variable interactions. Notably, rs5744657 (*POLK*) and rs4151276 (*MNAT1*) interacted with all 11 variables, followed by the exonic SNP rs3218786 (*POLI*), which interacted with sex, history of childhood sunburns, sunlamp use, time outdoors in the winter and age.

In the M-inv, we identified 29 significantly interacting SNPs across 23 genes. Similar to BCC, sunlamp use was the most common SNP-interacting variable, represented by 27 SNPs, of which 23 interacted exclusively with this variable and 6 of which represented the only exonic markers. Further, three *FANCA* SNPs (rs9926296, rs3743860 and rs2376883) which were found to interact with nearly all variables in the BCC group were found to interact with every variable in the M-inv group, further underscoring the importance of *FANCA* in the pathogenesis of both BCC and M-inv. In the M-inv group, rs3092829 (*ATM*) was the only other SNP that interacted with greater than one variable, including sex, skin color, tanning behavior, sunlamp use, hours spent outdoors in winter and summer, age and use of sun protection. Among both melanoma groups, we identified five SNPs (rs111885773 [*ENDOV*], rs150393409 [*FAN1*], rs61753893 [*FANCM*], rs114554002 [*RAD50*]) that interacted exclusively with the sunlamp use variable, suggesting shared genetic and environmental risk factors for both diseases. Moreover, we identified distinct SNPs from two additional genes (*ALKBH3* and *FANCF*) which interacted with different variables in each melanoma group. While we did not identify any significantly interacting SNPs that were common across all four skin cancers, distinct SNPs from 4 genes (*FAM193A*, *FANCA*, *MGMT* and *RAD51B*) were found to interact with at least one demographic and environmental/behavioral variable in each disease group. Figure 2 is an overview of

SNP-environment interaction analysis in the M-inv group (overviews of this analysis for other disease groups are presented in Figures S1–3). Figure 3 highlights genes pertaining to SNPs that overlapped across disease groups that could be of value for future individual risk assessment for two or more skin cancers.

Analysis of simultaneous skin cancer diagnoses

To further assess the shared risk between skin cancer subtypes, we performed pairwise analysis within our cohort of individuals diagnoses with two or more skin cancers (Figure 4). Out of 1,288 such individuals, representing approximately 5% of the total disease cohort, we identified the greatest overlaps between BCC and cSCC (n=550), and BCC and M-inv (n=487), accounting for over 80% of simultaneous diagnoses. The remaining proportion was made up of overlaps between BCC and M-is (n=144), cSCC and M-inv (n=45), cSCC and M-is (n=10), as well as those of individuals with three simulatenous diagnoses, including BCC, SCC and M-inv (n=45), and BCC, cSCC and M-is (n=16). We identified no individuals with four simultaneous diagnoses, nor of simultaneous diagnoses of M-is and M-inv.

Discussion

By leveraging robust datasets from the UKBB, we performed disease-environment, disease-gene and gene-environment investigations, and identified key variables underlying the complex interplay between 190 DNA repair genes and eleven environmental and demographic factors in the pathophysiology of melanoma and keratinocyte carcinomas. The diagnostic accuracy of our disease cohort is supported through the selection of individuals on the basis of the validated ICD coding system and subsequently verified using histopathological analysis. The first two of these analyses enabled us to identify eleven SNPs in seven key genes and five participant-related factors that are significantly associated with BCC and other skin cancer types, respectively. Notably, it was through joint analysis, taking together genetic and participant-related factors, that we were able to expand significant findings to cumulatively include 147 SNPs across 84 genes interacting with all eleven participant factors across every disease group. Our findings point to overlapping but distinct factors that mediate risk of each cancer and cumulatively underscore the importance of integrating environmental and demographic factors into genetic analyses to draw meaningful conclusions about skin cancer risk prognosis.

We first conducted disease-environment analyses and showed that sex, skin and hair color, skin tanning behavior, and use of sun protection show the greatest associations with risk of all four cancers. This is consistent with previous findings as skin-associated DNA damage upon exposure to UV light is a major driver of most skin cancers, while hair color is a weak proxy for skin tone/genetic makeup, and is not directly related to this effect(28). Moreover, male sex was positively associated with BCC and particularly cSCC, consistent with previous findings, and negatively associated with M-inv and M-is, in contrast to the literature(29, 30). Surprisingly at first, frequent use of sunscreen was greatly associated with all cancers. This surprising association was reported in prior studies(31–33). This paradoxical finding was the increasing risk of skin cancers with increased sunscreen

use, which we posit can be explained by greater exposure to UV light and/or a lack of reapplication of sunscreen throughout the day, or due to increased use of sun protection following skin cancer diagnosis. Collectively, however, these findings demonstrate the importance of adequate and frequent sunscreen use and minimization of exposure to UV light, particularly in individuals with fair skin.

We subsequently investigated the association between genetic markers related to DNA repair mechanisms and skin cancer and identified eleven significant markers across BCC and M-inv. Three of these markers (rs9926296, rs3743860, rs2376883), located in the *FANCA* gene (collectively, introns 29 and 31), were associated with both cancers. Two of these SNPs (rs3743860 and rs9926296) have been shown to be associated with colorectal cancer and generalized vitiligo respectively, pointing to the potentially shared pathways in both cancer and skin disease(34, 35). Moreover, rs12046289 was shown to confer two-fold increased risk of cutaneous melanoma among individuals with a strong family history, while rs1800347 was significantly associated with high-risk non-BRCA mutant breast cancer in a French Canadian cohort(36, 37). Nine of these SNPs (rs9926296, rs3743860, rs2376883, rs12046289, rs1800347, rs62989960, rs4942486, rs28928581 and rs4902628) are found in intron regions, while the other two (rs4149909 and rs61748588) are in exon regions. Notably, the identified G allele of rs4149909 is associated with a non-synonymous, missense mutation that substitutes asparagine for serine in the EXO1 protein sequence, with potential to disrupt protein structure and confer dysregulated DNA repair function, and has been associated with keratinocyte cancers in a large European cohort study(38). Further, the A allele of rs6178588 is associated with a synonymous mutation; despite maintaining wild-type protein sequence. Synonymous mutations have been shown to have potential to impact protein function through changes to mRNA splicing, translation and protein folding, disruption of microRNA-mediated gene regulation, and formation of novel haplotypes with altered gene function. In a similar fashion, intronic SNPs have been shown impact gene expression through altered gene interactions with transcription factors and long non-coding RNAs, as well as influencing epigenetic gene regulation through changes to genomic imprinting and chromatin-DNA interactions(39, 40). Collectively, these findings suggest that the identified SNPs may both play a regulatory role and have a direct impact on protein-coding sequence in mediating skin cancer risk.

Combining all available data, we then conducted gene-environment analyses, which provide the most informative conclusions about the interplay between DNA repair genes and participant-linked factors in each disease group. In BCC, the strongest genetic effects were in the *FANCA* and *DCLRE1B* genes, while the strongest environmental effects were linked to the number of lifetime sunburns (Figure S1); the latter is consistent with epidemiological literature which highlights the effect of lifetime sunburns and genetic predisposition as major contributors to BCC development(41). Similarly, cSCC is mediated by interactions between the number of lifetime sunburns and, to a lesser degree, aging appearance as interacting variables with many loci (Figure S2). M-inv shows similar trends with the three *FANCA* markers significant for all demographic and environmental variables, with the most common environmental interacting variables being lifetime sunburns and sunlamp use; these findings further support the importance of *FANCA* in predisposing to these cancers (Table 2, Figure 2, Figure S1). By contrast, M-is shows fewer significant interactions between a small

subset of specific genes and various environmental factors, and instead appears to be to a significant degree driven by sunlamp use, an environmental factor that shows interactions with numerous genes (Figure S3). These findings are also largely consistent with previous literature identifying the most important risk factors for each skin cancer group, namely chronic lifetime sun exposure in cSCC, number of lifetime sunburns and genetics in BCC, and sunlamp as well as sunburns in both *in-situ* and invasive melanoma(42–46).

For the purposes of using GWAS to predict skin cancer risk, as now performed by 23andMe® and AncestryDNA® for a variety of other diseases, our most consistent finding across keratinocyte carcinoma and invasive melanoma populations is the association of loci in the *FANCA* gene with all types of skin cancer analyzed especially for Caucasian individuals who report sunlamp use and a history of childhood sunburns. *FANCA* codes for a subunit of a protein family involved in post-replication repair, and mutation in these genes have been associated with increased risk of numerous hereditary and sporadic cancers(47, 48). Although another *FANCA* SNP has been previously associated with overall survival of melanoma, we have not observed reports of this particular marker in the context of BCC, cSCC, and M-inv(49). Nevertheless, our findings and prior reports underscore the importance of this gene as an important mediator of age, sex and behavior in the development of both melanoma and keratinocyte carcinomas.

Our study has several limitations. Our genetic analyses were limited to markers (SNPs) associated with 190 DNA repair genes, which are known to have far-reaching interactions with other pathways that were not captured here. Also, our study design does not capture additional environmental factors that may be relevant to disease pathophysiology. In addition, UKBB data does not delineate factors associated with skin cancer and is known to suffer from some degree of selection bias(50). Moreover, given that approximately 5% of our total disease cohort is made up of individuals with more than skin cancer diagnosis, we acknowledge the introduction of some degree of bias and potential for confounding in our analyses. Finally, our behavioral findings are self-reported by individuals, where recall bias can play a significant role, and do not account for changes in behavior before and after disease diagnosis.

Our study highlights the importance of conducting dynamic investigations of genetic variation by incorporating demographic, environmental and behavioral factors to fully appreciate the complex pathophysiology of skin cancers. Our findings highlight the prognostic value of specific SNPs in the *FANCA* gene in Caucasian individuals who report frequent sunlamp use and a history of childhood sunburns. This data will require validation in other populations where GWAS data is available and can be correlated with specific exposures. Additional SNPs identified in this report may also have significant association with the development of common skin cancers. Our findings have laid the foundation for a precision medicine approach to risk assessment based on molecular markers in key DNA repair genes and individual factors. Given recent increases in highly affordable risk assessments based on GWAS/SNP analyses, this work underscores that risk stratification can be greatly improved when diseases are studied holistically.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

References

1. Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. JAMA Dermatol. 2015;151(10):1081–6. [PubMed: 25928283]

2. Conte S, Aldien AS, Jette S, LeBeau J, Alli S, Netchiporouk E, et al. Skin Cancer Prevention across the G7, Australia and New Zealand: A Review of Legislation and Guidelines. Curr Oncol. 2023;30(7):6019–40. [PubMed: 37489567]

3. Conte S, Ghazawi FM, Le M, Nedjar H, Alakel A, Lagace F, et al. Population-Based Study Detailing Cutaneous Melanoma Incidence and Mortality Trends in Canada. Front Med (Lausanne). 2022;9:830254. [PubMed: 35308490]

4. Ransohoff KJ, Jaju PD, Tang JY, Carbone M, Leachman S, Sarin KY. Familial skin cancer syndromes: Increased melanoma risk. J Am Acad Dermatol. 2016;74(3):423–34; quiz 35–6. [PubMed: 26892652]

5. Que SKT, Zwald FO, Schmults CD. Cutaneous squamous cell carcinoma: Incidence, risk factors, diagnosis, and staging. J Am Acad Dermatol. 2018;78(2):237–47. [PubMed: 29332704]

6. Marzuka AG, Book SE. Basal cell carcinoma: pathogenesis, epidemiology, clinical features, diagnosis, histopathology, and management. Yale J Biol Med. 2015;88(2):167–79. [PubMed: 26029015]

7. Lagacé F, Noorah BN, Conte S, Mija LA, Chang J, Cattelan L, et al. Assessing Skin Cancer Risk Factors, Sun Safety Behaviors and Melanoma Concern in Atlantic Canada: A Comprehensive Survey Study. 2023;15(15):3753.

8. Cleaver JE, Crowley E. UV damage, DNA repair and skin carcinogenesis. Front Biosci. 2002;7:d1024–43. [PubMed: 11897551]

9. Lefrancois P, Xie P, Gunn S, Gantchev J, Villarreal AM, Sasseville D, et al. In silico analyses of the tumor microenvironment highlight tumoral inflammation, a Th2 cytokine shift and a mesenchymal stem cell-like phenotype in advanced in basal cell carcinomas. J Cell Commun Signal. 2020;14(2):245–54. [PubMed: 32198729]

10. Litvinov IV, Xie P, Gunn S, Sasseville D, Lefrancois P. The transcriptional landscape analysis of basal cell carcinomas reveals novel signalling pathways and actionable targets. Life Sci Alliance. 2021;4(7).

11. Xie P, Lefrancois P, Sasseville D, Parmentier L, Litvinov IV. Analysis of multiple basal cell carcinomas (BCCs) arising in one individual highlights genetic tumor heterogeneity and identifies novel driver mutations. J Cell Commun Signal. 2022;16(4):633–5. [PubMed: 35414145]

12. Gantchev J, Messina-Pacheco J, Martinez Villarreal A, Ramchatesingh B, Lefrancois P, Xie P, et al. Ectopically Expressed Meiosis-Specific Cancer Testis Antigen HORMAD1 Promotes Genomic Instability in Squamous Cell Carcinomas. Cells. 2023;12(12).

13. Kong D, Giovanello KS, Wang Y, Lin W, Lee E, Fan Y, et al. Predicting Alzheimer's Disease Using Combined Imaging-Whole Genome SNP Data. J Alzheimers Dis. 2015;46(3):695–702. [PubMed: 25869783]

14. Diogo D, Tian C, Franklin CS, Alanne-Kinnunen M, March M, Spencer CCA, et al. Phenome-wide association studies across large population cohorts support drug target validation. Nature Communications. 2018;9(1):4285.

15. Gruendner J, Wolf N, Tögel L, Haller F, Prokosch HU, Christoph J. Integrating Genomics and Clinical Data for Statistical Analysis by Using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): System Design and Implementation. J Med Internet Res. 2020;22(10):e19879. [PubMed: 33026356]

16. Abbasi J. 23andMe Develops First Drug Compound Using Consumer Data. JAMA. 2020;323(10):916-.

17. Crawford DC, Sedor JR. Biobanks Linked to Electronic Health Records Accelerate Genomic Discovery. J Am Soc Nephrol. 2021;32(8):1828–9. [PubMed: 34244324]

18. Khan R, Mittelman D. Consumer genomics will change your life, whether you get tested or not. Genome Biol. 2018;19(1):120. [PubMed: 30124172]

19. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med. 2009;6(2):e22. [PubMed: 19192942]

20. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. [PubMed: 25826379]

21. Chae YK, Anker JF, Carneiro BA, Chandra S, Kaplan J, Kalyan A, et al. Genomic landscape of DNA repair genes in cancer. Oncotarget. 2016;7(17):23312–21. [PubMed: 27004405]

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. [PubMed: 17701901]

23. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. ed. New York: Springer; 2002.

24. Hauck WW, Donner A. Wald's Test as Applied to Hypotheses in Logit Analysis. Journal of the American Statistical Association. 1977;72(360a):851–3.

25. Almli LM, Duncan R, Feng H, Ghosh D, Binder EB, Bradley B, et al. Correcting Systematic Inflation in Genetic Association Tests That Consider Interaction Effects: Application to a Genome-wide Association Study of Posttraumatic Stress Disorder. JAMA Psychiatry. 2014;71(12):1392–9. [PubMed: 25354142]

26. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

27. Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, et al. GenVisR: Genomic Visualizations in R. Bioinformatics. 2016;32(19):3012–4. [PubMed: 27288499]

28. Narayanan DL, Saladi RN, Fox JL. Ultraviolet radiation and skin cancer. Int J Dermatol. 2010;49(9):978–86. [PubMed: 20883261]

29. Bassukas ID, Tatsioni A. Male Sex is an Inherent Risk Factor for Basal Cell Carcinoma. J Skin Cancer. 2019;2019:8304271. [PubMed: 31772775]

30. Olsen CM, Thompson JF, Pandeya N, Whiteman DC. Evaluation of Sex-Specific Incidence of Melanoma. JAMA Dermatol. 2020;156(5):553–60. [PubMed: 32211827]

31. Wolf P, Quehenberger F, Mullegger R, Stranz B, Kerl H. Phenotypic markers, sunlight-related factors and sunscreen use in patients with cutaneous melanoma: an Austrian case-control study. Melanoma Res. 1998;8(4):370–8. [PubMed: 9764814]

32. Whiteman DC, Valery P, McWhirter W, Green AC. Risk factors for childhood melanoma in Queensland, Australia. Int J Cancer. 1997;70(1):26–31. [PubMed: 8985086]

33. Rueegg CS, Stenehjem JS, Egger M, Ghiasvand R, Cho E, Lund E, et al. Challenges in assessing the sunscreen-melanoma association. Int J Cancer. 2019;144(11):2651–68. [PubMed: 30447006]

34. Pardini B, Corrado A, Paolicchi E, Cugliari G, Berndt SI, Bezieau S, et al. DNA repair and cancer in colon and rectum: Novel players in genetic susceptibility. Int J Cancer. 2020;146(2):363–72. [PubMed: 31209889]

35. Jin Y, Birlea SA, Fain PR, Ferrara TM, Ben S, Riccardi SL, et al. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. Nat Genet. 2012;44(6):676–80. [PubMed: 22561518]

36. Liang XS, Pfeiffer RM, Wheeler W, Maeder D, Burdette L, Yeager M, et al. Genetic variants in DNA repair genes and the risk of cutaneous malignant melanoma in melanoma-prone families with/without CDKN2A mutations. Int J Cancer. 2012;130(9):2062–6. [PubMed: 21671477]

37. Litim N, Labrie Y, Desjardins S, Ouellette G, Plourde K, Belleau P, et al. Polymorphic variations in the FANCA gene in high-risk non-BRCA1/2 breast cancer individuals from the French Canadian population. Mol Oncol. 2013;7(1):85–100. [PubMed: 23021409]

38. Liyanage UE, Law MH, Han X, An J, Ong JS, Gharahkhani P, et al. Combined analysis of keratinocyte cancers identifies novel genome-wide loci. Hum Mol Genet. 2019;28(18):3148–60. [PubMed: 31174203]

39. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine Nicholas R, Chelala C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Research. 2020;48(W1):W185–W92. [PubMed: 32496546]

40. Deng N, Zhou H, Fan H, Yuan Y. Single nucleotide polymorphisms and cancer susceptibility. Oncotarget. 2017;8(66).

41. Armstrong BK, Kricker A, English DR. Sun exposure and skin cancer. Australas J Dermatol. 1997;38 Suppl 1:S1–6. [PubMed: 10994463]

42. Kilgour JM, Jia JL, Sarin KY. Review of the Molecular Genetics of Basal Cell Carcinoma; Inherited Susceptibility, Somatic Mutations, and Targeted Therapeutics. Cancers (Basel). 2021;13(15).

43. Conforti C, Zalaudek I. Epidemiology and Risk Factors of Melanoma: A Review. Dermatol Pract Concept. 2021;11(Suppl 1):e2021161S. [PubMed: 34447610]

44. Ghazawi FM, Cyr J, Darwich R, Le M, Rahme E, Moreau L, et al. Cutaneous malignant melanoma incidence and mortality trends in Canada: A comprehensive population-based study. J Am Acad Dermatol. 2019;80(2):448–59. [PubMed: 30092328]

45. Ghazawi FM, Le M, Lagace F, Cyr J, Alghazawi N, Zubarev A, et al. Incidence, Mortality, and Spatiotemporal Distribution of Cutaneous Malignant Melanoma Cases Across Canada. J Cutan Med Surg. 2019;23(4):394–412. [PubMed: 31132871]

46. Ghazawi FM, Lu J, Savin E, Zubarev A, Chauvin P, Sasseville D, et al. Epidemiology and Patient Distribution of Oral Cavity and Oropharyngeal SCC in Canada. J Cutan Med Surg. 2020;24(4):340–9. [PubMed: 32238063]

47. Del Valle J, Rofes P, Moreno-Cabrera JM, López-Dóriga A, Belhadj S, Vargas-Parra G, et al. Exploring the Role of Mutations in Fanconi Anemia Genes in Hereditary Cancer Patients. Cancers (Basel). 2020;12(4).

48. Chen H, Zhang S, Wu Z. Fanconi anemia pathway defects in inherited and sporadic cancers. Translational Pediatrics. 2014;3(4):300–4. [PubMed: 26835350]

49. Yin J, Liu H, Liu Z, Wang LE, Chen WV, Zhu D, et al. Genetic variants in fanconi anemia pathway genes BRCA2 and FANCA predict melanoma survival. J Invest Dermatol. 2015;135(2):542–50. [PubMed: 25243787]

50. Swanson JM. The UK Biobank and selection bias. Lancet. 2012;380(9837):110.
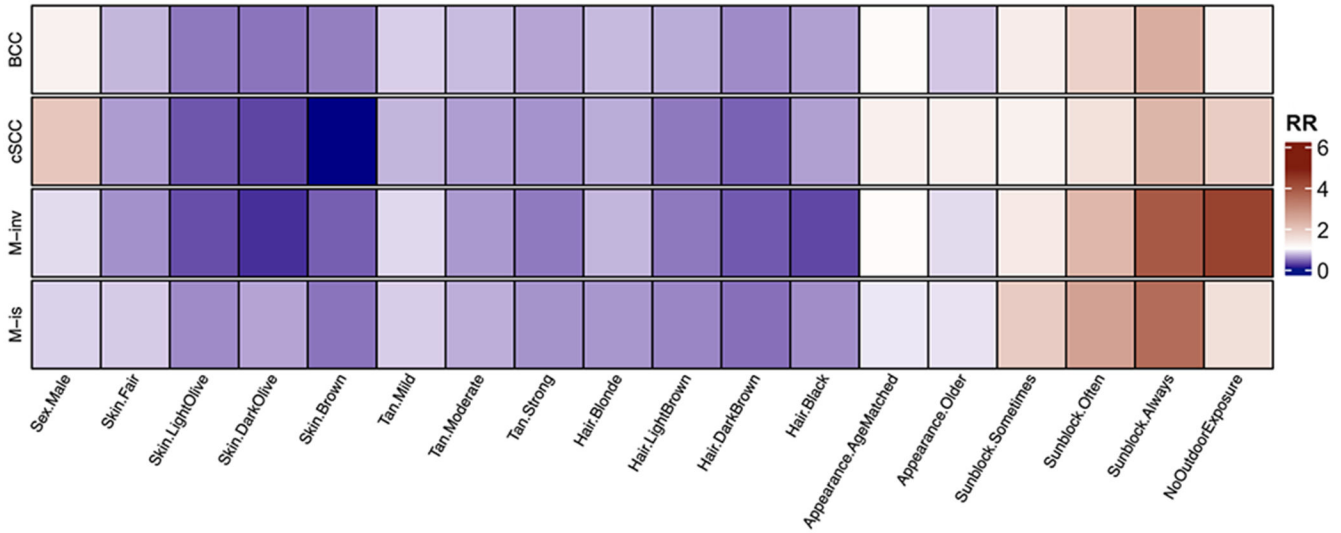
Figure 1.

Impact of demographic, environmental and behavioral variables on keratinocyte and melanoma skin cancer risk determined using multinomial logistic regression.

Darker skin and hair tones, as well as easy tanning ability correspond to a strongly decreased risk of all skin cancers. Age-matched and older appearance correspond to a mildly decreased risk of basal cell carcinoma (BCC), melanoma *in-situ* (M-is) and invasive melanoma (M-inv), and a slightly increased risk of cutaneous squamous cell carcinoma (cSCC). Male sex corresponds to moderately decreased risk of M-inv and M-is, moderately increased risk of cSCC and mildly increased risk of BCC. Frequent use of sun protection and limited outdoor exposure show moderately-to-strongly increased risk of all skin cancers.

Relative risk (RR) ratios represent disease risk; shades of red denote increased disease risk (RR > 1), while shades of blue denote decreased disease risk (RR < 1); white represents no difference in risk compared to healthy control (RR ~ 1). RR for each variable category is compared to a reference category (Sex – female; skin color – very fair; tan – never tan; hair color – red; appearance – young than age; use of sun protection – never/rarely). Significance ($p < 0.05$) was determined by the Wald Z test.
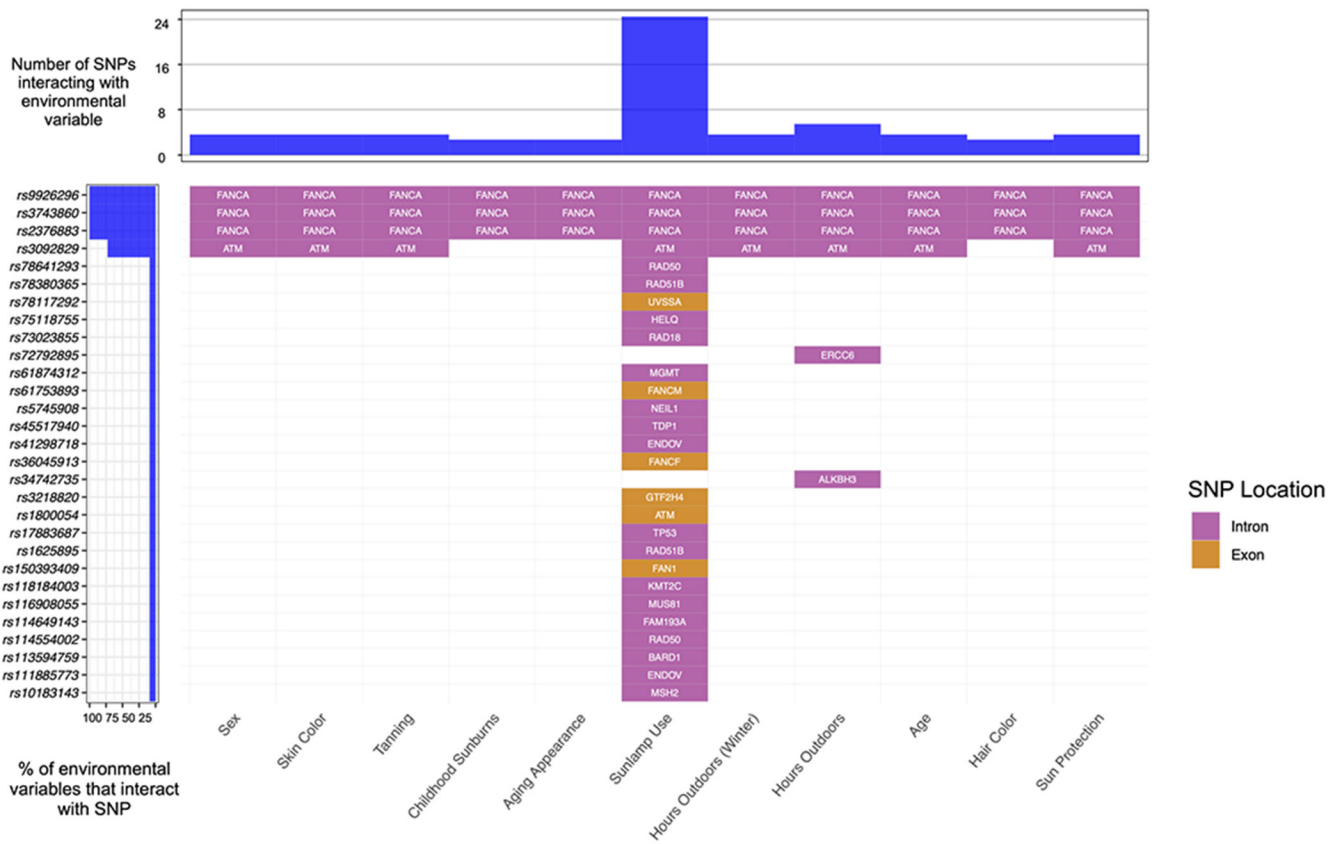
Figure 2.
Waterfall plot highlighting SNP-environment interactions in invasive melanoma (M-inv). Relationships were computed using a robust joint interaction test and assessed for significance using the False Discovery Rate (FDR) method. These findings highlight sunlamp use as a major interacting factor with DNA repair genes to mediate the risk of M-inv. This figure also highlights *FANCA* and *ATM* as critical genes which interact with nearly every demographic, environmental and behavioral variable.
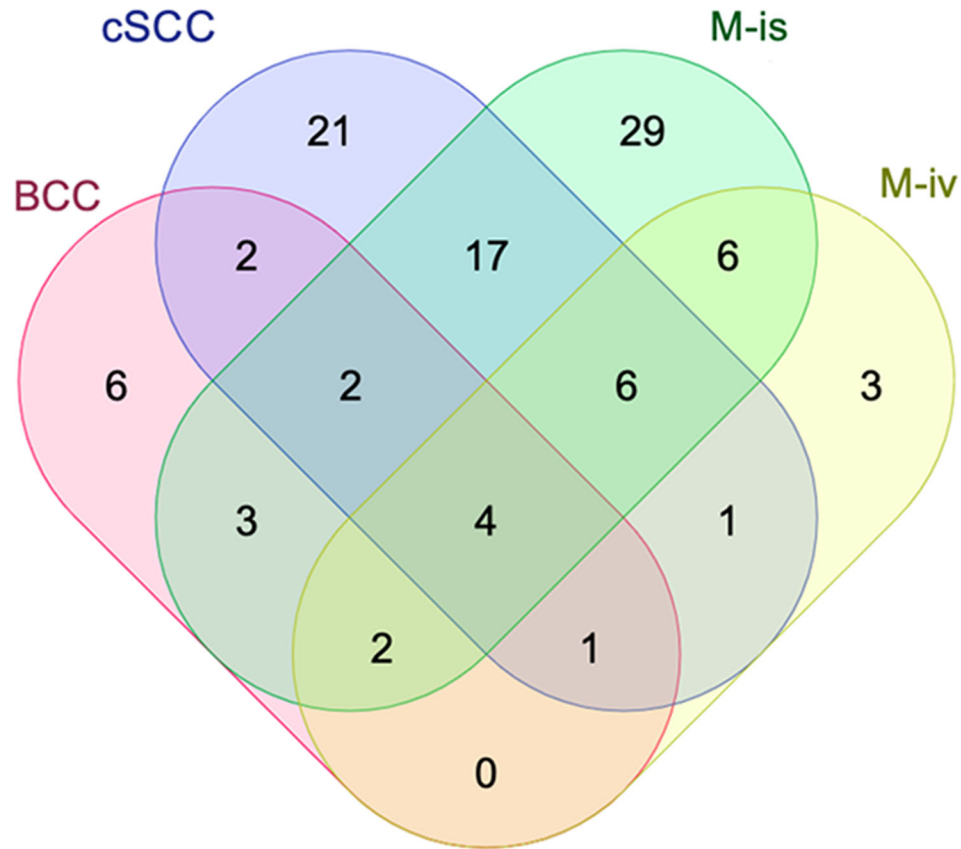
All Cancers	Keratinocyte carcinomas (KC)	Melanoma (invasive and in-situ)
FAM193A FANCA MGMT RAD51B	MNAT RECQL	ALKBH3 ENDOV FAN1 FANCF FANCM UVSSA

Figure 3.
Overview of DNA repair-associated genes that overlap between skin cancer groups. Findings are based on overlaps of significant ($q<0.05$) SNPs that interact with demographic, environmental and behavioral factors, as determined by a robust joint interaction test. These findings highlight the distinct, but shared pathogenesis of skin cancers and underscore key DNA repair-associated genes that are implicated in keratinocyte carcinomas, melanomas and all skin cancer groups.

	cSCC (657)	M-is (170)	M-inv (568)
BCC (1233)	602 0.4674	160 0.1287	523 0.4092
cSCC (657)		26 0.0325	81 0.0708
M-is (170)			0 0.0000

0   Jaccard index   1
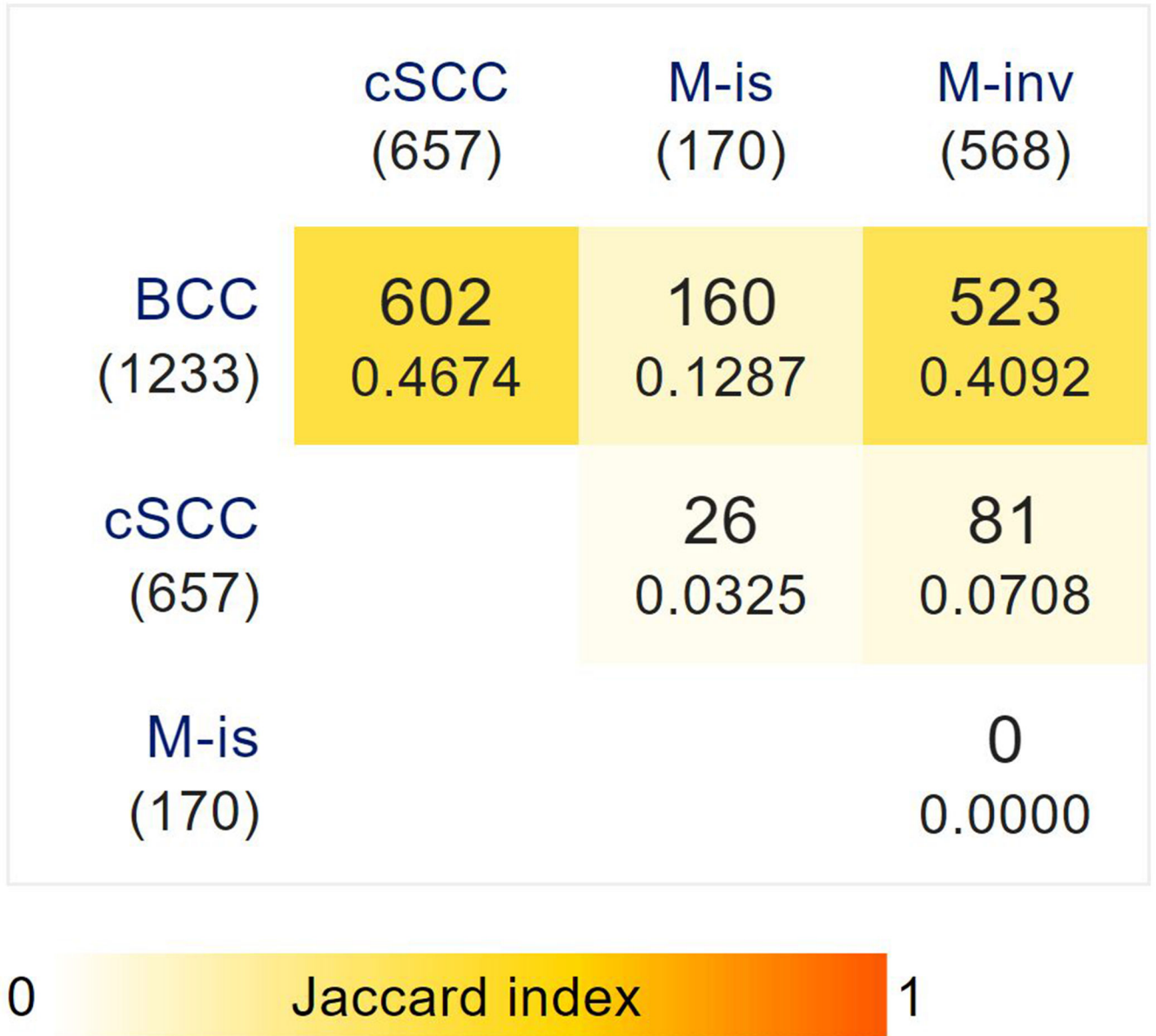
Figure 4.
Pairwise analysis of individuals with multiple skin cancer diagnoses.
Bracketed values denote the sample size of individuals in each skin cancer group from a total of 1,288 individuals with two or more simultaneous skin cancer diagnoses. Overlapping diagnoses are represented numerically and by pairwise similarity based on the Jaccard coefficient, and also incorporate individuals with three diagnoses. These findings represent the strong overlap between basal cell carcinoma (BCC) and cutaneous squamous cell carcinoma (cSCC), as well as between BCC and invasive melanoma (M-inv).

Table 1.

Overview of participant demographic, environmental and behavioral variables for healthy control and four skin cancer groups (basal cell carcinoma [BCC], cutaneous squamous cell carcinoma [cSCC], melanoma *in-situ* [M-is], and invasive melanoma [M-inv]). Data is based on available self-reported participant responses.

	Control		BCC		cSCC		M-is		M-inv	
n	448163		17222		2331		1158		3798	
Age, mean (SD)	57.2	8.2	59.2	8.9	63.3	7.5	58.5	9.5	54.9	11.2
Sex, female n (%)	244871	54.6	8520	49.5	884	37.9	686	59.2	2207	58.1
Hair color, n (%)										
Red	19593	4.4	1155	6.7	204	8.8	96	8.3	351	9.4
Blonde	49609	11.1	2177	12.7	352	15.1	146	12.6	642	17.2
Light brown	180633	40.4	7259	42.2	902	38.7	472	40.9	1541	41.1
Dark brown	170926	38.2	5570	32.4	683	29.3	372	32.2	1059	28.3
Black	20980	4.7	781	4.5	139	6.0	58	5.0	101	2.7
Other	5592	1.3	242	1.4	49	2.1	11	1.0	51	1.4
Skin color, n (%)										
Very fair	35238	8.0	1948	11.4	313	13.6	119	10.4	544	14.5
Fair	314221	71.0	12641	74.2	1726	74.9	854	74.3	2798	74.4
Light olive	84022	19.0	2216	13.0	249	10.8	157	13.7	390	10.4
Dark olive	7260	1.6	186	1.1	17	0.7	16	1.4	19	0.5
Brown	1948	0.4	55	0.3	0	0.0	3	0.3	11	0.3
Subjective age appearance, n (%)										
Younger than expected for age	302215	73.6	11341	72.5	1430	68.8	803	75.4	2519	72.5
As expected for age	100190	24.4	4051	25.9	599	28.8	242	22.7	895	25.8
Older than expected for age	8380	2.0	249	1.6	51	2.5	20	1.9	61	1.8
Ease of skin tanning, n (%)										
Never tan/only burn	75938	17.4	3712	22.1	562	24.7	262	23.2	918	24.7
Mildly or occasionally tanned	93066	21.3	3723	22.1	497	21.9	261	23.1	964	26.0
Moderately tanned	176432	40.4	6461	38.4	818	36.0	420	37.2	1294	34.9
Get very tanned	91493	20.9	2937	17.5	397	17.5	186	16.5	536	14.4
Use of sun/UV protection, n (%)										
Never/rarely	38197	8.6	871	5.1	129	5.6	41	3.6	147	3.9
Sometimes	149362	33.6	4474	26.2	625	27.1	300	26.0	784	20.8
Most of the time	162317	36.5	6607	38.6	831	36.0	453	39.2	1400	37.1
Always	92911	20.9	5083	29.7	708	30.7	357	30.9	1402	37.2
Do not go out into the sunshine	2404	0.5	68	0.4	15	0.7	4	0.4	40	1.1

	Control		BCC		cSCC		M-is		M-inv	
Annual frequency of solarium/sunlamp visits, mean (SD)	0.5	4.3	0.3	3.0	4.3	2.4	4.2	0.4	0.3	2.3
Number of childhood sunburns, mean (SD)	1.6	3.7	2.5	3.0	3.1	1.1	5.6	2.8	2.6	4.5
Daily hours spent outdoors, mean (SD)	3.8	2.4	4.1	2.3	2.1	1.8	2.3	3.9	3.8	2.3
Daily hours spent outdoors in winter, mean (SD)	1.9	1.9	2.0	1.7	2.2	1.8	1.7	1.9	1.9	1.8

Table 2.

Summary of significant SNP-disease interactions for invasive melanoma (M-inv; top panel) and basal cell carcinoma (BCC; bottom panel).

Chr	SNP	Allele	Odds Ratio	P-value (FDR)	Gene	Location
16	rs9926296	A	0.809446	3.77E-12	*FANCA*	intron 31
16	rs3743860	C	1.1965	1.29E-08	*FANCA*	intron 31
16	rs2376883	A	0.848934	1.92E-06	*FANCA*	intron 29
Chr	**SNP**	**Allele**	**Odds Ratio**	**P-value (FDR)**	**Gene**	**Location**
1	rs12046289	A	0.914388	0.000104	*DCLRE1B*	intron 2
1	rs4149909	G	1.12746	0.021555	*EXO1*	exon 7
6	rs61748588	A	1.18706	0.049773	*GTF2H4*	exon 5
8	rs28928581	C	1.08347	0.043767	*RRM2B*	intron 5
13	rs4942486	T	1.0483	0.011646	*BRCA2*	intron 21
14	rs4902628	C	0.958423	0.046463	*RAD51B*	intron 11
16	rs9926296	A	0.920691	2.42E-09	*FANCA*	intron 31
16	rs3743860	C	1.07056	5.88E-06	*FANCA*	intron 31
16	rs2376883	A	0.939527	0.000104	*FANCA*	intron 29
16	rs1800347	C	1.11993	0.004734	*FANCA*	intron 33
16	rs62989960	T	0.874634	0.004734	*FANCA*	intron 25

Odds ratios represent the probability of the tested minor allele of each SNP to be found in the disease versus healthy control group. Significance for each allele has been determined using the False Discovery Rate (FDR) method. Each SNP is annotated for its chromosomal position (Chr), the localizing gene, and its location (intron or exon) within the gene.