A Comprehensive Genetic Approach for Improving Prediction of Skin Cancer Risk in Humans

Ana I. Vazquez§*, Gustavo de los Campos§, Yann C. Klimentidis§, Guilherme J.M. Rosa§§,
Daniel Gianola§§ Nengjun Yi§ and David B. Allison§

§Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, AL, 35294
§§ Department of Animal Sciences, University of Wisconsin-Madison, WI, 53705

Reference numbers:

dbGaP Study Accession: phs000007.v1.p1
IRB protocol number: X100712003,
IRB protocol title: “Prediction of Diseases Phenotypes Using Whole Genome Markers from the Framingham Population Data Set”.

Copyright 2012.
Running head:

WHOLE GENOME PREDICTION OF SKIN CANCER

Key words:

Skin Cancer Risk
Whole Genome Prediction
Prediction of Complex Traits and Diseases
Pedigree Predictions
Genomic Predictions

To whom correspondence should be addressed

Ana I. Vazquez, PhD.
Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham,
1665 University Boulevard, Ryals Public Health Building 414, Birmingham, AL 35294
AL, 35294
e-mail: anainesvs@gmail.com
Phone: (205) 975 9268
Fax: (205) 975-2541
ABSTRACT

Prediction of genetic risk for disease is needed for preventive and personalized medicine. Genome wide association studies have found unprecedented numbers of variants associated with complex human traits and diseases. However, these variants explain only a small proportion of genetic risk. Mounting evidence suggests that many traits, relevant to public health, are affected by large numbers of small-effect genes and that prediction of genetic risk to those traits and diseases could be improved by incorporating large numbers of markers into whole-genome prediction (WGP) models. We developed a WGP model incorporating thousands of markers for prediction of skin cancer risk in humans. We also considered other ways of incorporating genetic information into prediction models, such as family history or ancestry (using principal components, PC, of informative markers). Prediction accuracy was evaluated using the area under the receiver operating characteristic curve (AUC) estimated in a cross-validation. Incorporation of genetic information (i.e., familial relationships, PC or WGP) yielded a significant increase in prediction accuracy: from an AUC of 0.53 for a baseline model that accounted for non-genetic covariates to AUCs of 0.58 (pedigree), 0.62 (PC), and 0.64 (WGP). In summary, prediction of skin cancer risk could be improved by considering genetic information and using a large number of SNPs in a WGP model, which allows for the detection of patterns of genetic risk that are above and beyond those that can be captured using family history. We discuss avenues for improving prediction accuracy and speculate on the possible use of WGP to prospectively identify individuals at high risk.

INTRODUCTION
Skin cancer is the most common form of cancer, and its incidence has increased in recent decades. In Queensland, Australia, it has been estimated that half of the population is likely to develop skin cancer during their lifetime ("World Cancer Report" 2008). In its most severe form (i.e., melanoma), skin cancer can be deadly. Though protection against sunburn (ZIEGLER et al. 1994) is widely believed to reduce the harmful effects of sun exposure on the skin, many individuals continue to seek out intense sun exposure without such protection (ROBINSON 1990). This may be due in part to belief that their risk of skin cancer is too low to be a serious concern. Such beliefs can be maintained and rationalized by the observation that many individuals exposed to such risk do not experience the adverse event. If individuals could be provided with personalized information about their individual risk, it might promote greater use of preventive measures among those at greatest risk.

Although ultraviolet (UV) exposure and light skin pigmentation are major risk factors for all types of skin cancers (e.g., it is estimated that 80% of melanoma is caused by ultraviolet damage to sensitive skin, (IARC 1992)), evidence suggests that genetic factors can also play a role, independent of skin pigmentation. Predictive models are usually based on standard covariables and family history. Additionally, several genetic variants, such as the MC1R, ASIP, TYR, EXOC2, UBAC2 and the 1p36 and 1q42 loci have been shown to be associated with basal and squamous cell carcinomas, as well as with melanomas, independent of skin pigmentation (GUDBJARTSSON et al. 2008; STACEY et al. 2008). These variants typically account for a small proportion of genetic-based disease risk (HAN et al. 2006; PHAROAH 2008).
As with other phenotypic and disease traits, the inability of loci discovered by genome-wide association studies (GWAS) to explain a substantial proportion of heritability has led to much debate regarding where this so-called ‘missing heritability’ lies (Manolio et al. 2009). It has been suggested that the underlying genetic architecture of many human traits and diseases may involve a substantial number of small-effect genes, thus conforming to the so-called infinitesimal model of quantitative genetics (Fisher and others 1918; Bulmer 1980; Lander and Schork 1994; Goddard and Hayes 2007). However, in most genetic risk prediction models currently being tested, only a few (i.e. less than 500 SNPs) statistically significant SNPs are included.

The recognition that complex human traits and diseases could be affected by a large number of genes has motivated many researchers in other fields (Wray et al. 2008; Lee et al. 2008; Purcell et al. 2009; Hill 2010; Yang et al. 2010; de los Campos et al. 2010a) to propose the use of statistical methods tailored for the prediction of complex traits. These methods, largely developed in the field of animal breeding, were first proposed by Meuwissen et al. (Meuwissen et al. 2001), who suggested predicting genetic factors by regressing phenotypes on a large number of markers covering the entire genome. The markers are assumed to be in linkage disequilibrium (LD) with one or many loci affecting the phenotypic traits, and the estimates of individual marker effects are expected to be small (Goddard and Hayes 2007). Such models and variations thereof have been used successfully in animal and plant breeding for prediction of production-related traits (de los Campos et al. 2009; Hayes et al. 2009; VanRADEN et al. 2009; Crossa et al. 2010; Weigel et al. 2010). More recently, several authors have proposed and used this methodology for the prediction of complex human traits such as height (Yang et al. 2010; Makowsky et al. 2011) and several cancer outcomes (Vazquez 2010).
In this study, we determine whether genetic predisposition to skin cancer could be used to predict disease outcome. Compared with height, skin cancer is less heritable, more complex, and highly relevant. Due to these features, we compared different methods to account for genetic susceptibility to skin cancer: (1) pedigree-based and SNP-based predictions via (2) WGP for liability to skin cancer using thousands of evenly distributed markers across the genome and via (3) the principal components of a subset of independent SNPs (previously used to predict geographical origin). To perform this study, we extended the Bayesian LASSO (Park and Casella 2008) regression with a probit link (Dempster and Lerner 1950) to model skin cancer.

**MATERIALS AND METHODS**

**Data**

The data set consists of 5,132 participants from the Framingham Heart Study, which has collected phenotypic information across three generations of families (Dawber et al. 1951, 1963). Subjects in this study have been characterized every other year from adulthood to death on risk factors, outcomes of physical exams, and disease status. Participants included in our study belong to the original cohort (n=1,498) and to the offspring cohort (n=3,634), with a total of 2,319 males and 2,813 females, all of whom were genotyped. Subjects from the Third Generation cohort were not included in our study because the follow-up period of this cohort was too short. The skin cancer outcomes were collected by Bernard E. Kreger (Boston University, Boston, MA, USA,
study accession number pht000039) (KREGER et al. 1991). The study declares cancerous all subjects whose pathology reports confirm their cancer. After the 1980’s, cases were validated using medical records. The available data represent primary tumors only. The skin cancer outcomes study was updated in 2006, containing life-long follow up, i.e., 1948 to 2006 for the Original cohort and 1971 to 2006 years for the Offspring cohort.

All subjects were genotyped for single nucleotide polymorphisms (SNPs) with the Affymetrix 500K chip. Due to computational limitations (memory requirements), we fitted models using 41,188 evenly spaced SNPs. Evidence from US Holstein cattle has indicated that predictive ability for several traits does not increase markedly when using over 10K SNPs for a panel of 50K maximum (VAZQUEZ et al. 2010). Nevertheless, the degree of linkage disequilibrium differs across species. However, a recent study on human height with the same data set indicated that, in this population and with this sample size (n=5,117), the increase in predictive ability when using more than 30K SNPs was limited (MAKOWSKY et al. 2011).

**Statistical Models**

**Full Model.** The outcome \( y_i \) was defined as presence \( y_i = 1 \) or absence \( y_i = 0 \) of skin cancer at any site, excluding skin of labia majora, vulva, penis and scrotum, as a primary tumor site. We modeled probability of skin cancer using the probit link or threshold model (DEMPSTER and LERNER 1950; HARVILLE and MEE 1984). Here, probability of disease equals the standard normal cumulative density function, \( \Phi() \), evaluated at a subject specific risk score, with either model:
\[ \eta_i = \beta_0 + \sum_{j=1}^{p_1} x_{1ij} \beta_{1j} + \sum_{j=1}^{p_2} x_{2ij} \beta_{2j}, \text{ or} \]

\[ \eta_i = \beta_0 + \sum_{j=1}^{p_1} x_{1ij} \beta_{1j} + u_i, \]

which was represented as the sum of an intercept \((\beta_0)\) plus a regression on the ‘fixed effects’ of sex (as dummy variable), cohort (a factor with three levels), and ethnicity covariates (explained below) \(\sum_{j=1}^{p_1} x_{1ij} \beta_{1j}\), plus either a ‘random effects’ regression on marker genotypes \(\sum_{j=1}^{p_2} x_{2ij} \beta_{2j}\) or a ‘random effects’ \(u_i\) being the genetic liability to skin cancer derived based on pedigree connections. Therefore, the joint conditional probability of the data, \(y = \{y_i\}\), given the unknown regression coefficients, \(\beta = \{\beta_0, \beta_{11}, \ldots, \beta_{1p_1}, \beta_{21}, \ldots, \beta_{2p_2}\}\) and \(u = \{u_1, \ldots, u_{5,132}\}\), was:

\[
p(y|\beta, u) = \prod_{i=1}^{5,132} \left\{ \Phi(\eta_i) \right\}^{y_i} \left[ 1 - \Phi(\eta_i) \right]^{1-y_i} \}
\]

Assigning a prior density to the vector of model unknowns, \(\beta_0, \beta, u\), completes the Bayesian model. We assigned a flat prior to the intercept and to the effects of sex, cohort, and ethnicity covariates. This yielded estimates of effects comparable to those obtained with maximum likelihood, and used the Bayesian Lasso of Park and Casella (PARK and CASELLA 2008) to structure the prior density of marker effects. This prior density yielded shrunken estimates of marker effects and has been successfully used for WGP (DE LOS CAMPOS et al. 2009). The joint prior density was:

\[
p(\beta_0, \beta_1, \beta_2, u, \tau^2, \lambda) \propto \left[ \prod_{j=1}^{p_1} N(\beta_{1j}, 0, \tau_j^2) \right] \times \left[ \prod_{j=1}^{p_2} \text{Exp}(\tau_j^2 | \lambda^2) \right] \times \text{G}(\lambda^2 | \alpha_1, \alpha_2) \times \text{N}(u | 0, \Lambda \sigma_u^2) \times \chi^{-2}(\sigma_u^2 | S, df),
\]
where $N(\beta_j, 0, \tau_j^2)$ is a normal density assigned to the $j^{th}$ marker effect, $j \in \{1, \ldots, p\}$ with prior mean equal to zero and marker specific prior variance $(\tau_j^2)$, $\text{Exp}(\tau_j^2 | \lambda^2)$ an exponential prior for variances of marker effects $\tau_j^2$, $G(\lambda^2 | \alpha_1, \alpha_2)$ a gamma distribution for $\lambda^2$ with shape and rate parameters $\alpha_1$ and $\alpha_2$, respectively, $N(0, A \sigma_u^2)$ a normal distribution for $u$ with mean 0 and variance $\sigma_u^2$ times $A$, which is the additive genetic relationship matrix based on the pedigree, and finally $\chi^2(\sigma_u^2) | S, df$ an Inverse Chi-squared distribution for $\sigma_u^2$ with parameters $S$ and $df$. In the analysis, these hyper-parameter values were set to $\alpha_1 = 1.5$ and $\alpha_2 = 10^{-4}$; this gives a relatively flat prior over a wide range of values of the regularization parameter (see Pérez et al. (2010) for further details), as $S$ and $df$ were 0.19 and 5, respectively. Models were fitted using a modified version of the BLR package (DE LOS CAMPOS and Pérez 2010) of R (R DEVELOPMENT CORE TEAM 2010), which can be used for regressions for binary outcomes according to the model described above.

**Sequence of models.** Using the specification described above, we defined a sequence of models and evaluated the performance of each model using cross-validation. Our baseline model (Covariates) included only the effects of sex and cohort. This model was first extended by adding to the regression the effects of the first two principal components of a set of 1,000 European ethnicity informative SNPs previously reported (model denoted as PC-SNP). The panel of ethnicity informative SNPs used here were those reported by Drineas et al. (DRINEAS et al. 2010). Figure 1 shows a scree plot of the first 20 eigenvalues derived from the markers reported by Drineas et al. In a number of studies, the first two principal components (PC1 and PC2) of a large set of genetic variants have been shown to be effective predictors of the ancestral/geographical origin (lati-
tude and longitude) of individuals of European descent (Price et al. 2008; Tian et al. 2008; Novembre et al. 2008; Drineas et al. 2010). For this reason we included 2 PC in the model, even though PC2 is only slightly higher than following PCs. Additionally, the Covariates model was extended by adding subsets of SNPs (250, 500, 1K, 5K, 10K and 41K, K=thousand) distributed over the whole human genome; these models are denoted as .25K-SNP, .5K-SNP, 1K-SNP, 5K-SNP, 10K-SNP, and 41K-SNP, respectively. The 41K SNPs were obtained by choosing approximately one of every 12 SNPs from the original SNP panel. The SNPs in the smaller sets are all included in the larger sets. In this series of increasingly denser models, we aim to discover how many markers are needed to increase the predictive ability. Finally, we extended the Covariates model by adding a random effect representing a regression on the pedigree, and this model was denoted as Pedigree.

*Estimated probabilities and odds ratio.* Probabilities and odds ratio for the relative risk of developing skin carcinoma for groups of sex, cohort, and ethnicity were estimated with the PC-SNP model at a fixed level of the dichotomous co variables and fixing the principal components at the mean, first, and third percentile value of the eigenvectors of PC1 and PC2 (see Table 2). Similarly, probabilities and odds ratio of the genetic effects \( \sum_{j=1}^{p_j} x_{2j} \beta_{2j} \) were estimated with the 41K-SNP model for differences in the first and third percentile of the genetic effects for fixed levels of sex and cohort.

*Assessment of model prediction performance.* Models were compared based on prediction accuracy, evaluated in a 20-fold cross-validation with subjects assigned to folds at random. The 20-fold cross-validation yielded predictions of risk scores \( \{\hat{\eta}_i\} \) which were derived without using
the $i^{th}$ observation or any of the observation assigned to the same fold to which the $i^{th}$ observation was assigned. Using the pairs of points $\{y_i, \hat{y}_i\}$, we estimated false positive rate and Area Under the Receiver Operating Characteristic Curve (AUC, see (FAWCETT 2006)) using the R package ROCR (SING et al. 2005).

To assess the uncertainty of the AUC estimate due to sampling variability, we performed 500 random partitions in training and testing sets with sizes equal to those of the cross-validation folds (testing in n=257 subjects, and training n=4,875 subjects), maintaining the sizes of the sub-sets in the cross-validation (i.e., on each replicate, 5% of the individuals were randomly assigned to testing and the remaining 95% for training). Each replicate yielded an estimate of AUC by model, and variability across replicates was reflective of uncertainty due to sampling of training and testing data sets. From this analysis, we reported the number of times one model outperformed another for AUC.

**RESULTS**

**Descriptive Statistics and Parameter Estimates**

The skin cancer incidence in our dataset was 14.1% for the entire period evaluated, starting at 1948 with the Original cohort and at 1971 with the Offspring cohort, and followed until the 2006 update. Incidence did vary, however, across sex and cohorts. The incidence was higher in males (16%) than in females (13%), and higher in the Original cohort (17%), which had a longer follow-up period, than in the Offspring cohort (13%). The first two eigenvectors of the principal
component decomposition (PCs) derived from 1,000 ethnicity-informative SNP are displayed in Figure 2, panel a. Panel b shows the empirical distribution of PC1 for individuals with and without skin cancer. The PC1 has the highest discriminating power (Figure 2 panel a) and, at the marginal level, lower values at the eigenvector were associated with higher incidence of skin cancer (Figure 2 panel b). This PC1 has been reported to track Northern vs. Southern European ancestry (Drineas et al. 2010). Further evidence of the marginal association of incidence of skin cancer and the PC1 is given in Table 1, where average incidence of skin cancer is presented by quartiles of PC1 and PC2.

In the PC-SNP model, the effects of cohort, sex and the first 2 PCs are estimated jointly. We found that the estimated coefficient for male sex in the liability scale was 0.18 [0.09, 0.26] (posterior mean and 95% credibility region in brackets). This estimate implies an increased higher risk of developing skin cancer in males, relative to females. The estimated coefficient for the Original cohort with respect to Offspring cohort as baseline was 0.32 [0.23, 0.42], also in liability scale. This implies higher risk of developing skin cancer for members of the Original cohort and likely reflects the effect of a longer follow-up period for members of this cohort (Original cohort started 23 years before the Offspring cohort). The estimated coefficients for the first and second PCs were -11.19 [-14.47, -7.87] for the PC1 and 12.97 [9.41, 16.62] for PC2, indicating that risk increases as PC1 decrease and as PC2 increase. All 95% confidence regions for the estimated effects did not include zero, showing evidence of non-null effects of the predictors considered on skin cancer risk. All the estimates presented above are in the scale of the linear predictor (or liability scale). Given the non-linearity of the model, these results are difficult to interpret. Table 2 shows the estimated probability of skin cancer risk (and estimates of 95% posterior credi-
ibility regions) for different combinations of the predictor variables using the PC-SNP model. The 95% credibility regions of the probability estimates for the two cohorts and for gender do not overlap, suggesting significant differences for these predictors. The odds ratio for the cohort variable is 1.76 [1.50, 2.10] (Original relative to Offspring cohort) in males, and 1.81 [1.52, 2.18] in females, while the odds ratio for sex is 1.36 [1.17, 1.58] (male relative to female) in the Original cohort and 1.39 [1.19, 1.64] in the Offspring cohort. All were estimated at the mean value of the 2 PCs.

When the Covariates model was extended by adding genetic effects connected by the pedigree, the predicted genetic effects \( \hat{u}_i \) in the liability scale ranged between -.5 to 1.37 (Figure 3). Likewise, the Covariates model was also extended by adding the joint effects of SNPs evenly spaced along the genome (from 250 to 41K). In models including genome-wide SNPs, the total contribution of all SNPs, to the cancer risk, is summarized by the linear score \( \hat{g}_i = \sum_{j=1}^{p_2} x_{2j} \hat{\beta}_{2j} \), where \( x_{2j} \) are marker genotypes and \( \hat{\beta}_{2j} \) are estimates of marker effects. In our sample, this score (\( \hat{g}_i \)) ranged between -1.07 to 2.7 for the model with 41K markers (Figure 3). These results suggest the existence of variation due to genetic factors which was captured by either the pedigree or the markers. Figure 3 shows the predicted genetic scores derived from the pedigree \( \hat{u}_i \) and the WGR (41KSNPs) \( \hat{g}_i \) (both derived from models fitted to the entire dataset). The correlation between the genetic scores was 0.783. Both scores exhibit a bi-modal distribution: the group with lower scores corresponds to individuals with no personal or family history of skin cancer, while the second group primarily includes individuals with some personal or family history of skin cancer. The within-group dispersion of the pedigree-based score around the mean of
the clusters is much smaller than that of the score derived from the WGR. This occurs because, although WGR captures family history, this approach also allows for the borrowing of information across nominally-unrelated individuals.

**Evaluation of the Models Predictive Performances**

The estimates presented in the preceding section indicate that all the predictor variables are significantly associated with the risk of developing skin cancer. Additionally, we evaluated the prediction accuracy of each of the models with a 20-fold cross-validation to assess how useful each of these models is in the assessment of risk in individuals with yet-to-be-observed skin cancer outcomes.

Figure 4 shows the AUC obtained in the 20-fold cross-validation for (a) the Covariates, Pedigree, PC-SNP, and 41KSNP, and for (b) genomic-enabled models at increasing SNP density from zero to 41K. The covariates model had an AUC of 0.534 (baseline model). Accounting also for the pedigree relations yielded an AUC of 0.579, improving the AUC by 8.4% (calculated as 100 × (0.579 − 0.534)/0.534). The PC-SNP had an AUC of 0.622, 16.5% higher than that of the baseline model. Finally, the 41K-SNP model had an AUC of 0.635, 18.9% higher than that of the baseline model. The evaluation of the AUC of the models including sex, cohort, and varying numbers of evenly spaced SNP showed a monotonic increase in AUC with the number of SNPs, from the covariates model (with zero SNPs) to 41K-SNP (Figure 4, panel b).
Family relationships between training and testing sets has been shown to affect prediction accuracy (Habier et al. 2010; Pérez-Cabal et al. 2012). To assess the impact of family relationships on the prediction accuracy derived from models incorporating pedigree or markers, we calculated the AUC of each of the models for subjects that did not have data from relatives in the training data sets (n=871) and for those that had at least one relative in the training data set (n=4,248) (see Table 3). Results show that in the Pedigree and the 41K-SNPs model, part of the gains in prediction accuracy can be explained by information provided by relatives. However, as shown in Figure 4: (a) the 41K-SNP model outperformed the pedigree model, and (b) the 41K-SNP model had higher predictive accuracy than all the other models for individuals whose risk was predicted without having any relative in the training data set (1st row in Table 3). Therefore, we conclude that although part of the prediction accuracy of the 41K-SNP model can be explained by information coming from relatives, this model is capturing patterns of genetic risk beyond those that can be captured by family history.

Assessing the sampling variability of the AUC estimates. The above results suggest that whole-genome markers can increase prediction accuracy of skin-cancer susceptibility by a non-negligible margin. To evaluate uncertainty about these point-estimates, we replicated a training-testing evaluation 500 times (see methods section above). The Covariates model was improved by the Pedigree model in 70% of the replicates, by the PC-SNP model in 90% of the replicates, and by the 41K-SNP model in 94% of the replicates. The 41K-SNP model had higher prediction accuracy than the Pedigree model in 90% of the replicates, and was higher than the PC-SNP model in 66% of the replicates. Figure 5 shows predictive correlation and AUC results by replicate for two models simultaneously (y-axis and x-axis). At the 45 degree line, both models per-
formed equally, while above the line, the model represented on the y-axis performed better, and *vice versa*.

**DISCUSSION**

*Factors affecting skin cancer.* Skin cancer is the most frequent form of cancer. Further, non-melanoma skin cancer is the most frequent type of cancer in light-skinned populations (“World Cancer Report” 2008). In our sample, skin cancer was also the most prevalent type of cancer. The main risk factors for skin carcinogenesis are ultraviolet light exposure, skin type, and geographical location (e.g., no-melanoma skin cancer is 5 times higher in the USA and 20 to 40 times higher in Australia than in Europe (“World Cancer Report” 2008)). In our study, geographical location was not a source of variation, since all data came from the same geographical location (Framingham, MA). However, the harm from sun exposure may have increased over time due to, e.g., atmospheric changes. Indeed, it has been estimated that incidence of non-melanoma skin cancer increased by 77% from 1992 to 2006 (STERN 2010 p. 13). The estimated effect of cohort in our data might be simultaneously reflecting two factors which may have opposing effects: subjects from the Original cohort have had longer exposure (Offspring cohort data collection started 23 years later); and year of exposure (since skin cancer incidence has increased over the past 50 years). Gender differences have also been reported in the literature, and they indicate a lower incidence among females (DIEPGEN and MAHLER 2002). Our results are consistent with this (Table 2). The association between sex and risk of developing skin cancer has been largely attributed to different lifestyles (e.g., males are more exposed to sun and less likely to use sun protection) (MCCARTHY et al. 1999). Additionally, there is evidence of sex-based-biological differences at the skin level, relevant to skin cancer liability (THOMAS-AHNER et al. 2007).
**Prediction of genetic risk to skin cancer.** Previous studies (HAN et al. 2006; “World Cancer Report” 2008; GUDBJARTSSON et al. 2008; STACEY et al. 2008) have indicated that genetic factors play an important role in predisposition to skin cancer. However, predictive models for skin cancer, although accurate, do not usually account for genetic factors (e.g., SOONG et al. 2010). In this study, we show that considering genetic information, under the form of familial relationships, SNP-derived PCs or WGP using markers evenly-distributed in the genome can increase the prediction accuracy of risk of developing skin cancer.

Simply considering family history, under the form of pedigree-relationships linked to phenotypes, increased prediction accuracy. However, there is a limit to how much prediction accuracy can be gained by considering family history alone. One of the limitations of using pedigree connections in a predictive model is that family size in humans is usually small. Our sample, for instance, has not related individuals. Other limitations of models using pedigrees are (1) capturing important elements of genetic variability, such as variability due to substructure or admixture, and (2) these models cannot describe genetic differences between individuals with identical pedigree (e.g., full sibs) due to sampling of alleles at the meiosis. Therefore, describing genetic background using markers can potentially boost prediction accuracy above and beyond what can be achieved using family history. In agreement with previous studies in animals (VANRADEN et al. 2009), our study confirms this and suggests that simply considering 2-SNP-derived PCs can increase prediction accuracy substantially. Skin color, and therefore ethnicity, is known to be highly correlated with skin cancer (“World Cancer Report” 2008), and skin color varies even among individuals of European descent. We found that cancer incidence was high at low levels of PC1 and at high levels of PC2. In previous studies, PC1 among Europeans has been found to correspond to ancestry along the Northwest to Southeast European geographical axis (CAMPBELL et al. 2005; NOVEMBRE et al. 2008; DRINEAS et al. 2010). Therefore, one possible explanation of the increase in prediction accuracy obtained by considering 2-SNP derived PCs is that these PCs are capturing ancestry, which correlates with skin color and with risk of developing skin cancer.
Naturally, there is also a limit in the proportion of genetic variability at causal loci that can be explained by 2 PCs. Our study confirms this, the WGP model using 41K-SNP outperformed all the other models we considered, including the one with PCs. The prediction accuracy of WGP increases monotonically with marker density, a finding that is consistent with the hypothesis that genetic risk to skin cancer is affected by a large number of variants. A polygenic genetic architecture has also been suggested for other human traits (e.g., VATTIKUTI et al. 2012).

Empirical evidence for complex traits in animals (VAZQUEZ et al. 2010) and humans (MAKOWSKY et al. 2011) has shown that prediction accuracy increases with marker density, and our results are consistent with this. However, prediction accuracy depends on many other factors, perhaps most importantly on the size of the training set (\(n\)) (VANRADEn et al. 2009). Preliminary evidence of (un-published) studies we conducted with human height suggests that the level at which the curve relating prediction accuracy and marker density reaches a plateau is highly dependent on marker density. Prediction accuracy also depends on the selection criteria of the markers incorporated as predictors in the model (VAZQUEZ et al. 2010). In that study, the predictive correlation obtained with a set of 300 markers highly associated with the trait of interest was, in average for 6 traits in cattle, 0.18 higher than that obtained with a set of 300 evenly spaced markers. Here, markers are evenly distributed, representing the whole genome, and are not particularly associated with skin cancer. We expect predictive accuracy to increase if the markers were selected based on their association with the disease. Therefore, we speculate that further increases in marker density, targeting markers associated with skin cancer, accompanied by increases in sample size, could increase prediction accuracy of WGP even further.

The WGR models implemented in this study account for additive effects of markers. Potentially, these additive models could be extended to account for interactions of alleles within loci (i.e., dominance) and between loci (i.e., epitasis). With \(p\) markers, modeling additive and dominance
effects involves estimating $2p$ effects, and this can be done by using the methods similar to those described in this article. However, modeling epistatic interactions is much more difficult because the number of contrasts required and, consequently, the number of parameters to be estimated, grows exponentially with the number of markers and the order of the interaction. Alternatively, one can attempt to capture departures from the linear model using WGP with non-parametric procedures, such as penalized Neural Networks or Reproducing Kernel Hilbert Spaces (Gianola et al. 2006; De los Campos et al. 2010b). However, even in cases where complex interactions among alleles hold at the causal level, a large proportion of inter-individual differences in genetic risk may manifest as additive variance (Hill et al. 2008), and the information provided by data for estimation of non-additive effects may be small (Hill et al. 2008). Because of this, it is not necessarily the case that use of models that account for non-additive effects will yield higher prediction accuracy than that of an additive model.

**Summary.** Although accurate in predicting survival rate once the signs of the disease are present, previous predictive models for skin cancer do not account for genetic susceptibility factors (Soong et al. 2010), and therefore, they have limited use for preventive measure that can be applied early in life. In our study, prediction substantially improved by using genetic parameters in the predictive models. Further, methods including genome-wide markers information outperformed models with genetic risk estimates derived from the pedigree. WGP is a promising tool for estimating individual genetic predisposition to skin cancer before it is detected or even developed. We speculate that genomic information may be used to prospectively identify individuals with particularly high risk of developing skin cancer.
ACKNOWLEDGEMENTS

The authors gratefully acknowledge scientists conducting the Framingham Heart Study for collecting, analyzing, and providing the genotypic and phenotypic data used in the present study. We also acknowledge the invaluable dedication of all participants in the Framingham Heart Study for their contribution over many years. The authors also acknowledge NIH grants R01GM077490, R01GM099992, R01GM101219 the Kraft grant, as well as the Wisconsin Agriculture Experiment Station for their financial support.

REFERENCES


DAWBER T. R., MEADORS G. F., MOORE JR F. E., 1951  Epidemiological approaches to heart

DE LOS CAMPOS G., GIANOLA D., ALLISON D. B., 2010a  Predicting genetic predisposition in

DE LOS CAMPOS G., GIANOLA D., ROSA G. J. M., WEIGEL K. A., CROSSA J., 2010b  Semi-
parametric genomic-enabled prediction of genetic values using reproducing kernel Hil-

DE LOS CAMPOS G., NAYA H., GIANOLA D., CROSSA J., LEGARRA A., MANFREDI E., WEIGEL K.,
COTES J. M., 2009  Predicting quantitative traits with regression models for dense mo-
lecular markers and pedigree. Genetics 182: 375.


DIEPGEN T., MAHLER V., 2002  The epidemiology of skin cancer. British Journal of Dermatolo-

ogy 146: 1–6.

DRINEAS P., LEWIS J., PASCHOU P., 2010  Inferring geographic coordinates of origin for Euro-


FISHER R. A., OTHERS, 1918  The correlation between relatives on the supposition of Mendelian


IARC, 1992  IARC monographs on the evaluation of carcinogenic risks to humans.


PÉREZ P., DE LOS CAMPOS G. DE LOS, CROSSA J., GIANOLA D., 2010  Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. The Plant Genome Journal 3: 106.


**TABLES AND FIGURES**

**Table 1.** Incidence of skin cancer by levels defined using the first and second eigenvectors of the ethnicity SNP-derived principal components.

<table>
<thead>
<tr>
<th>Group</th>
<th>( v_i \leq q_{.25} )</th>
<th>( q_{.25} &lt; v_i \leq q_{.50} )</th>
<th>( q_{.50} &lt; v_i \leq q_{.75} )</th>
<th>( v_i &gt; q_{.75} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(^{st}) Principal Component</td>
<td>0.178</td>
<td>0.150</td>
<td>0.126</td>
<td>0.098</td>
</tr>
<tr>
<td>2(^{nd}) Principal Component</td>
<td>0.100</td>
<td>0.150</td>
<td>0.152</td>
<td>0.163</td>
</tr>
</tbody>
</table>

\( v_i \): eigenvector of the principal component (1\(^{st}\) or 2\(^{nd}\)); \( q \): corresponding quartile.

**Table 2:** Estimated probabilities of developing skin cancer for different levels of the predictor variables, derived from a model including sex, cohort and the first 2 principal components of 1000-ethnicity informative SNPs.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Sex</th>
<th>Probability of developing skin cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimate</td>
</tr>
<tr>
<td>Original</td>
<td>Male</td>
<td>0.242</td>
</tr>
<tr>
<td>Offspring</td>
<td>Male</td>
<td>0.153</td>
</tr>
<tr>
<td>Original</td>
<td>Female</td>
<td>0.190</td>
</tr>
<tr>
<td>Offspring</td>
<td>Female</td>
<td>0.115</td>
</tr>
</tbody>
</table>

**Table 3:** Area Under the Curve estimated in the subjects that have no relatives in the training set and in the subjects that do, for all the models.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Pedigree</th>
<th>PC-SNP</th>
<th>41K-SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>No relatives in training set</td>
<td>0.540</td>
<td>0.549</td>
<td>0.635</td>
</tr>
<tr>
<td>At least 1 relative in training set</td>
<td>0.531</td>
<td>0.583</td>
<td>0.619</td>
</tr>
</tbody>
</table>
Figure 1: First 20 eigenvalues derived from ethnicity-informative panel of one thousand SNPs.
Figure 2: (a) First (x-axis) and second (y-axis) principal components eigenvectors derived from 1,000 ethnicity-informative SNPs (red dots correspond to subjects that developed skin cancer, grey dots correspond to healthy subjects); (b) Empirical distribution of the first principal component separated by cancerous or healthy subjects.
Figure 3: Scatter plot of the pedigree-based predicted genetic risk to skin cancer, and the SNP-based ones ($\hat{u}_i$ and $\hat{g}_i$, respectively), as well as the histogram of their distribution.
Figure 4: Mean area under the curve for the 20 folds of the cross-validation for (a) a model without any genetic information, and two models with genetic information, one including pedigree and a WGP model, and (b) for WGP models of increasing number of SNPs.
**Figure 5**: AUC in 500 random training - testing sets of genetically informed models (Pedigree model, PC-SNP and 41K-SNP models) vs. the baseline model (Covariates), and average AUC for the 500 training –testing in 41K-SNP model vs. Pedigree model.