Genetics, Caffeine Consumption, Height and Non-Melanoma Skin Cancer

Genetics, caffeine consumption, height and non-melanoma skin cancer

Xin Li

A Dissertation Submitted to the Faculty of

The Harvard T.H. Chan School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Science

in the Department of Epidemiology

Harvard University

Boston, Massachusetts.

May, 2016

Dissertation Advisor: Dr. Jiali Han Xin Li

Genetics, caffeine consumption, height and non-melanoma skin cancer

Abstract

Non-melanoma skin cancer (NMSC), including basal and squamous cell carcinoma (BCC and SCC, respectively), is the most common malignancy among populations of European ancestry. It is estimated that over 2 million cases of NMSC occur each year in the United States, with the incidence continues to increase. This disease imposes a growing burden on healthcare system, making it an important public health issue. However, understanding of its etiology and biological mechanisms remains incomplete.

In Chapter 1 of this dissertation, we applied a novel approach that integrates skin expression-related single-nucleotide polymorphisms (eSNPs) and pathway analysis to identify potential novel biological pathways that are associated with BCC risk. We evaluated the associations of skin eSNPs with BCC among 2,323 cases and 7,275 controls of European ancestry, and assigned them to the pathways defined by KEGG, GO, and BioCarta databases. Three KEGG pathways (colorectal cancer, regulation of actin cytoskeleton, and basal cell carcinoma) and two GO pathways (cellular component disassembly involved in apoptosis, and nucleus organization) showed significant association with BCC risk. Our results indicate that genes that are undetectable by conventional genome-wide association studies (GWASs) are

significantly associated with risk of BCC as groups.

In Chapter 2, we tested gene-caffeine consumption interaction on BCC risk in a genome-wide analysis. We determined that SNP rs142310826 shows a genome-wide significant interaction with caffeine consumption ($p = 1.78\times10^{-8}$ for interaction, $p = 0.64$ for heterogeneity between genders) on BCC risk. We also found several loci that modify the caffeine-BCC association differently in men and women. This study is proof of concept that inclusion of environmental factors can help identify genes that are missed in conventional GWASs.

In Chapter 3, we prospectively examined the risk of SCC and BCC in relation to adult height. After controlling for potential confounding factors, the hazard ratios were 1.09 (95% CI: 1.03, 1.16) and 1.10 (95% CI: 1.07, 1.12) for the associations between every 10cm increase in height and risk of SCC and BCC respectively. However, no significant association was observed between height-related SNPs and risk of these diseases.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the kind help and support of many individuals. I would like to extend my sincere thanks to all of them, though only some of the names are mentioned here.

First and foremost, I would like to express my gratitude to my advisor, Dr. Jiali Han, for his guidance, encouragement and patience over the past years. Dr. Han led me into the field of genetic epidemiology and imparted his knowledge and expertise to me generously. His advice and support has been invaluable to me because it not only make me a better researcher, but also a better person in everyday life, for which I am extremely grateful. I thank him for being such a great mentor and friend.

I would also like to thank my committee members, Dr. Immaculata De Vivo, Dr. Edward Giovannucci, and Dr. Liming Liang, for their thoughtful and constructive comments on my dissertation. I really enjoy every discussion we had, and feel so lucky and honored to have them in my research committee.

In addition, I would like to acknowledge my group members and my fellow graduate students for their assistance and encouragement. I am also indebted to my respected teachers and other members of the Department of Epidemiology and Harvard T.H. Chan School of Public Health.

Finally, I must thank all my family members. They have always been there for me and I am thankful for everything they have helped me achieve.

INTRODUCTION

Non-melanoma skin cancer (NMSC) is the most common malignancy among populations of European ancestry [1]. NMSC usually refers to basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), which together account for more than 95% of all NMSC cases [1]. BCC arises from the pluripotential primordial cells in the basal layer of the epidermis [2]. Though the tumor tends to grow slowly and rarely metastasizes to other organs or causes death, it can lead to extensive tissue destruction, resulting in considerable morbidity if treated insufficiently [2]. SCC is a malignant proliferation of the keratinocytes in the epidermis or its dermal appendages [3]. Unlike BCC, cutaneous SCC is more likely to invade other tissues and can be fatal [2]. Non-melanoma skin tumors also arise from other cell types of skin, such as lymphocytes, vascular endothelial cells, and Merkel cells [1, 4]. However, these forms of NMSC are so rare compared to BCC and SCC that they will not be discussed in this dissertation.

The incidence of BCC and SCC is difficult to determine, because neither is registered by most cancer surveillance systems. The majority of data on NMSC incidence come from local studies in specific geographical locations [5]. Using national population-based data sources, Rogers et al. [6] reported that 2,152,500 persons were treated for NMSC in 2006 in the United States. Those authors also estimated a 4.2% annual average increase in NMSC cases among the Medicare population from 1992 to 2006 [6]. Increases in NMSC incidence have also been documented in other countries [5, 7-9]. Such trends may be linked to increased sun-seeking behaviors, use of artificial UV tanning beds, ozone depletion, increases in life expectancy, more

exposures to chemical carcinogens, and improvements in public awareness or medical detection [10, 11]. Although the incidence of NMSC increases with age [2], it is now becoming more common among younger people, probably because they spend more time in the sun with their skin exposed. The risk of BCC and SCC is higher in men, but incidence in women has been steadily increasing [12].

Exposure to ultraviolet (UV) radiation is thought to be the main cause of NMSC, though its effects on BCC and SCC are different. A strong dose-response association has been found between cumulative lifetime sun exposure and SCC, whereas sun exposure during childhood and adolescence appear to be more important for development of BCC [13, 14]. Compared to a similar degree of continuous exposure, intense intermittent exposure to the sun is associated with a higher risk of BCC [13, 15]. A survey conducted in eight geographically diverse locations in the United States showed an inverse association between latitude and risk of NMSC (i.e., the farther from the equator/higher the latitude, the lower the risk) [16], providing further evidence for the important role of UV radiation in the development of NMSC. Physical characteristics such as fair complexion, red/blonde hair color, and lower tanning ability are also risk factors for NMSC [17].

Researchers have also investigated the potential roles of other environmental risk factors. Previous studies generally support a positive association between smoking and risk of SCC [18-20]. In the Nurses' Health Study (NHS), smokers had a 50% higher risk of SCC than non-smokers [18]. However, such an association has not been found in the majority of studies of smoking and BCC risk [21-24]. The association between alcohol intake and risk of NMSC may

vary for different types of alcoholic beverages. Some studies found BCC to be positively associated with total alcohol intake and white wine consumption, but inversely associated with red wine consumption [24-26]. The association between alcohol intake and risk of SCC has been sparsely reported. Occupational [27, 28] and therapeutic [29-31] ionizing radiation has been reported to increase BCC and SCC risk. Atomic bomb survivors were found to be more likely to develop BCC after long latent periods; however SCC risk was not increased [32]. Dietary factors with antioxidant and anti-inflammatory effect have also been hypothesized to modify risk of skin cancer. Previous animal studies and epidemiological studies have shown that caffeine administration/consumption is associated with lower risk of skin cancers [33-38]. Using data from the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS), Song *et al.* reported that increased caffeine intake is significantly associated with reduced risk of BCC [39]. Anthropometric indicators such as height and BMI are also thought to affect skin cancer risk through modification of metabolism and/or immune function, or simply through the association of body size with number of target cells. BMI appeared to be inversed associated with development of NMSC in the NHS and HPFS [40]. However, the relationship between height and NMSC has not been specifically investigated by cohort studies.

Genetic factors have also been implicated as playing critical roles in the development of NMSC [41]. In a large twin cohort study, the heritability of non-melanoma skin cancer was estimated to be 43% (95%CI: 26% - 59%) [42]. Mutations in the patched 1 gene (*PTCH1*), a tumor suppressor in the hedgehog signaling pathway, have been found in 30%-40% of sporadic BCC cases [43, 44]. *RAS* mutations [45-48] and UV-induced somatic p53 mutations [49-53] have

also been described in NMSC, though the reported rates varied among studies. Most recently, genome-wide association studies (GWAS) have identified a number of genetic loci associated with risk of NMSC. A meta-analysis of previous GWAS results showed that the significant loci are 1p36, 1q42, 5p13, 5p15, 7q32, 9q21, 11q14-21, 12q11-13, 16q24, and 20q11.2-12 for BCC, and 5p13 for SCC [54]. Despite recent advances, understanding of the biological mechanisms underlying these complex diseases remains incomplete.

Although GWAS have revolutionized our ability to identify disease susceptibility loci or markers associated with them, they do have limitations [55]. On the one hand, most common DNA variants with moderate effect size have not yet been identified by GWAS because of a lack of power [56]. Realizing this, new approaches are emerging to enhance the information extracted from current GWAS data. These include association analyses using multiple genetic markers [57-59], association tests with imputed genotypes [60, 61], association analyses incorporating linkage information [62], and more recently pathway-based association approaches [63]. On the other hand, the majority of GWAS to date have tested for association only between individual genetic markers and traits of interest without taking interactions into consideration. As a result, they may have failed to discover loci that influence disease only in the presence of particular genetic or environmental exposure [64]. It has been widely accepted that "gene-environment (G-E) interaction" is ubiquitous in the development of most complex diseases. Therefore, including key environmental factors in genetic association studies is anticipated to be an important next step for understanding the genetic structure of complex multifactorial disorders.

In this dissertation, we first conducted a pathway analysis of expression-related SNPs

(eSNPs) on risk of BCC. Then, we investigated the interaction between caffeine consumption and genetic markers in BCC risk using genome-wide data. Lastly, we comprehensively examined the association between height and risk of incident SCC and BCC using data from the NHS and the HPFS. These analyses may enhance our understanding of the etiology of NMSC and provide more insights into the biological mechanisms of these diseases.

References

1. Patel, R.S., et al., *Non-melanoma Skin Cancer.* Montgomery PQ, Evans PHR, Gullane PJ. Principles and Practice of Head and Neck Surgery and Oncology. 2nd. United kingdom: Informa healthcare, 2009: p. 480-95.
2. Tung, R.C. and A.T. Vidimos, *Non-melanoma skin cancer.* Retrieved March, 2002. **29**: p. 2007.
3. Schwartz, R.A., *Squamous cell carcinoma*, in *Skin Cancer*. 1988, Springer. p. 36-47.
4. Madan, V., J.T. Lear, and R.-M. Szeimies, *Non-melanoma skin cancer.* The Lancet, 2010. **375**(9715): p. 673-685.
5. Lomas, A., J. Leonardi‐Bee, and F. Bath‐Hextall, *A systematic review of worldwide incidence of nonmelanoma skin cancer.* British Journal of Dermatology, 2012. **166**(5): p. 1069-1080.
6. Rogers, H.W., et al., *Incidence estimate of nonmelanoma skin cancer in the United States, 2006.* Archives of dermatology, 2010. **146**(3): p. 283-287.
7. Demers, A.A., et al., *Trends of nonmelanoma skin cancer from 1960 through 2000 in a*

Canadian population. Journal of the American Academy of Dermatology, 2005. **53**(2): p. 320-328.

8. Staples, M.P., et al., *Non-melanoma skin cancer in Australia: the 2002 national survey and trends since 1985.* Medical Journal of Australia, 2006. **184**(1): p. 6.

9. Brewster, D., et al., *Recent trends in incidence of nonmelanoma skin cancers in the East of Scotland, 1992–2003.* British Journal of Dermatology, 2007. **156**(6): p. 1295-1300.

10. Diepgen, T. and V. Mahler, *The epidemiology of skin cancer.* British Journal of Dermatology, 2002. **146**(s61): p. 1-6.

11. Wu, S., et al., *Basal-cell carcinoma incidence and associated risk factors in US women and men.* American journal of epidemiology, 2013. **178**(6): p. 890-897.

12. Czarnecki, D., et al., *The changing face of skin cancer in Australia.* International journal of dermatology, 1991. **30**(10): p. 715-717.

13. Kricker, A., et al., *Does intermittent sun exposure cause basal cell carcinoma? a case‐control study in Western Australia.* International Journal of Cancer, 1995. **60**(4): p. 489-494.

14. Rosso, S., et al., *The multicentre south European study'Helios'. II: Different sun exposure patterns in the aetiology of basal cell and squamous cell carcinomas of the skin.* British journal of cancer, 1996. **73**(11): p. 1447.

15. Rubin, A.I., E.H. Chen, and D. Ratner, *Basal-cell carcinoma.* New England Journal of Medicine, 2005. **353**(21): p. 2262-2269.

16. Scotto, J., T.R. Fears, and J.F. Fraumeni, *Incidence of nonmelanoma skin cancer in the*

United States. 1983, US Department of Health and Human Services Washington.

17. Han, J., G.A. Colditz, and D.J. Hunter, *Risk factors for skin cancers: a nested case–control study within the Nurses' Health Study.* International journal of epidemiology, 2006. **35**(6): p. 1514-1521.

18. Grodstein, F., F.E. Speizer, and D.J. Hunter, *A prospective study of incident squamous cell carcinoma of the skin in the nurses' health study.* Journal of the National Cancer Institute, 1995. **87**(14): p. 1061-1066.

19. Moore, S.R., et al., *The epidemiology of lip cancer: a review of global incidence and aetiology.* Oral diseases, 1999. **5**(3): p. 185-195.

20. Karagas, M.R., et al., *Risk of subsequent basal cell carcinoma and squamous cell carcinoma of the skin among patients with prior skin cancer.* Jama, 1992. **267**(24): p. 3305-3310.

21. Hunter, D.J., et al., *Risk factors for basal cell carcinoma in a prospective cohort of women.* Annals of epidemiology, 1990. **1**(1): p. 13-23.

22. Green, A., et al., *Skin cancer in a subtropical Australian population: incidence and lack of association with occupation.* American Journal of Epidemiology, 1996. **144**(11): p. 1034-1040.

23. Lear, J., et al., *Risk factors for basal cell carcinoma in the UK: case-control study in 806 patients.* Journal of the Royal Society of Medicine, 1997. **90**(7): p. 371-374.

24. Freedman, D.M., et al., *Risk of basal cell carcinoma in relation to alcohol intake and smoking.* Cancer Epidemiology Biomarkers & Prevention, 2003. **12**(12): p. 1540-1543.

25. Fung, T.T., et al., *Intake of alcohol and alcoholic beverages and the risk of basal cell carcinoma of the skin.* Cancer Epidemiology Biomarkers & Prevention, 2002. **11**(10): p. 1119-1122.

26. Soleas, G.J., et al., *A comparison of the anticarcinogenic properties of four red wine polyphenols.* Clinical biochemistry, 2002. **35**(2): p. 119-124.

27. Wang, J.-X., et al., *Cancer among medical diagnostic x-ray workers in China.* Journal of the National Cancer Institute, 1988. **80**(5): p. 344-350.

28. Yoshinaga, S., et al., *Nonmelanoma skin cancer in relation to ionizing radiation exposure among US radiologic technologists.* International journal of cancer, 2005. **115**(5): p. 828-834.

29. Karagas, M.R., et al., *Risk of basal cell and squamous cell skin cancers after ionizing radiation therapy.* Journal of the National Cancer Institute, 1996. **88**(24): p. 1848-1853.

30. Levi, F., et al., *Skin cancer in survivors of childhood and adolescent cancer.* European Journal of Cancer, 2006. **42**(5): p. 656-659.

31. Johnson, T.M., et al., *Squamous cell carcinoma of the skin (excluding lip and oral mucosa).* Journal of the American Academy of Dermatology, 1992. **26**(3): p. 467-484.

32. Ron, E., et al., *Skin tumor risk among atomic-bomb survivors in Japan.* Cancer Causes & Control, 1998. **9**(4): p. 393-401.

33. Abel, E.L., et al., *Daily coffee consumption and prevalence of nonmelanoma skin cancer in Caucasian women.* european Journal of Cancer prevention, 2007. **16**(5): p. 446-452.

34. Rees, J.R., et al., *Tea consumption and basal cell and squamous cell skin cancer: results*

of a case-control study. Journal of the American Academy of Dermatology, 2007. **56**(5): p. 781-785.

35. Stensvold, M.I. and B.K. Jacobsen, *Coffee and cancer: a prospective study of 43,000 Norwegian men and women.* Cancer Causes & Control, 1994. **5**(5): p. 401-408.

36. Lu, Y.-P., et al., *Topical applications of caffeine or (−)-epigallocatechin gallate (EGCG) inhibit carcinogenesis and selectively increase apoptosis in UVB-induced skin tumors in mice.* Proceedings of the National Academy of Sciences, 2002. **99**(19): p. 12455-12460.

37. Kerzendorfer, C. and M. O'Driscoll, *UVB and caffeine: inhibiting the DNA damage response to protect against the adverse effects of UVB.* Journal of Investigative Dermatology, 2009. **129**(7): p. 1611-1613.

38. Lou, Y.-R., et al., *Effects of oral administration of tea, decaffeinated tea, and caffeine on the formation and growth of tumors in high-risk SKH-1 mice previously treated with ultraviolet B light.* Nutrition and cancer, 1999. **33**(2): p. 146-153.

39. Song, F., A.A. Qureshi, and J. Han, *Increased caffeine intake is associated with reduced risk of basal cell carcinoma of the skin.* Cancer research, 2012. **72**(13): p. 3282-3289.

40. Pothiawala, S., et al., *Obesity and the incidence of skin cancer in US Caucasians.* Cancer causes & control, 2012. **23**(5): p. 717-726.

41. Lear, J., et al., *Basal cell carcinoma: from host response and polymorphic variants to tumour suppressor genes.* Clinical and experimental dermatology, 2005. **30**(1): p. 49-55.

42. Mucci, L.A., et al., *Familial Risk and Heritability of Cancer Among Twins in Nordic Countries.* JAMA, 2016. **315**(1): p. 68-76.

43. Taipale, J. and P.A. Beachy, *The Hedgehog and Wnt signalling pathways in cancer.* nature, 2001. **411**(6835): p. 349-354.

44. Gailani, M.R., et al., *The role of the human homologue of Drosophila patched in sporadic basal cell carcinomas.* Nature genetics, 1996. **14**(1): p. 78-81.

45. van der Schroeff, J.G., et al., *Ras oncogene mutations in basal cell carcinomas and squamous cell carcinomas of human skin.* Journal of investigative dermatology, 1990. **94**(4): p. 423-425.

46. Lieu, F.-M., et al., *Low incidence of Ha-ras oncogene mutations in human epidermal tumors.* Cancer letters, 1991. **59**(3): p. 231-235.

47. Campbell, C., A. Quinn, and J. Rees, *Codon 12 Harvey‐ras mutations are rare events in non‐melanoma human skin cancer.* British Journal of Dermatology, 1993. **128**(2): p. 111-114.

48. Pierceall, W.E., et al., *Ras gene mutation and amplification in human nonmelanoma skin cancers.* Molecular carcinogenesis, 1991. **4**(3): p. 196-202.

49. Rady, P., et al., *p53 mutations in basal cell carcinomas.* Cancer research, 1992. **52**(13): p. 3804-3806.

50. Shea, C.R., et al., *Overexpression of p53 protein in basal cell carcinomas of human skin.* The American journal of pathology, 1992. **141**(1): p. 25.

51. Moles, J., et al., *p53 gene mutations in human epithelial skin cancers.* Oncogene, 1993. **8**(3): p. 583-588.

52. Konishi, K., et al., *Analysis of p53 gene mutations and loss of heterozygosity for loci on*

chromosome 9q in basal cell carcinoma. Cancer letters, 1994. **79**(1): p. 67-72.

53. Ziegler, A., et al., *Mutation hotspots due to sunlight in the p53 gene of nonmelanoma skin cancers.* Proceedings of the National Academy of Sciences, 1993. **90**(9): p. 4216-4220.

54. Gerstenblith, M.R., J. Shi, and M.T. Landi, *Genome‐wide association studies of pigmentation and skin cancer: a review and meta‐analysis.* Pigment cell & melanoma research, 2010. **23**(5): p. 587-606.

55. Erb, P., et al., *Apoptosis and pathogenesis of melanoma and nonmelanoma skin cancer*, in *Sunlight, Vitamin D and Skin Cancer*. 2008, Springer. p. 283-295.

56. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* science, 2008. **322**(5903): p. 881-888.

57. Li, M., et al., *ATOM: a powerful gene-based association test by combining optimally weighted markers.* Bioinformatics, 2009. **25**(4): p. 497-503.

58. Gauderman, W.J., et al., *Testing association between disease and multiple SNPs in a candidate gene.* Genetic epidemiology, 2007. **31**(5): p. 383-395.

59. Kwee, L.C., et al., *A powerful and flexible multilocus association test for quantitative traits.* The American Journal of Human Genetics, 2008. **82**(2): p. 386-397.

60. Li, Y., et al., *Genotype imputation.* Annual review of genomics and human genetics, 2009. **10**: p. 387.

61. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nature Reviews Genetics, 2010. **11**(7): p. 499-511.

62. Roeder, K., et al., *Using linkage genome scans to improve power of association in*

genome scans. The American Journal of Human Genetics, 2006. **78**(2): p. 243-252.

63. Wang, K., M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* The American Journal of Human Genetics, 2007. **81**(6): p. 1278-1283.

64. Kraft, P. and D.J. Hunter, *The challenge of assessing complex gene-environment and gene-gene interactions.* Human Genome Epidemiology, 2008: p. 165.

CHAPTER 1

Pathway analysis of expression-related SNPs on genome-wide association study of basal cell carcinoma

Xin Li[1], Liming Liang[1], Immaculata De Vivo[1,2], Jean Y. Tang[3] and Jiali Han[4]

[1] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

[2] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA;

[3] Department of Dermatology, Stanford University School of Medicine, Redwood City, CA, USA;

[4] Department of Epidemiology, Fairbanks School of Public Health, Indiana University, and Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, IN, USA

Abstract

Conventional genome-wide association studies (GWAS) have primarily focused on the association between individual genetic markers and risk of disease. We applied a novel approach that integrates skin expression-related single-nucleotide polymorphisms (eSNPs) and pathway analysis for GWAS of basal cell carcinoma (BCC) to identify potential novel biological pathways. In all, 70,932 eSNPs in skin tissue with significance levels of 10^{-5} were obtained from the Multiple Tissue Human Expression Resource (MuTHER). We evaluated the associations of these functionally annotated SNPs with BCC among 2,323 cases and 7,275 controls of European ancestry, and then assigned them to the pathways defined by KEGG, GO, and BioCarta databases. Three KEGG pathways (colorectal cancer, regulation of actin cytoskeleton, and basal cell carcinoma), two GO pathways (cellular component disassembly involved in apoptosis, and nucleus organization), and four BioCarta pathways (Ras signaling pathway, T cell receptor signaling pathway, Ras-independent pathway in natural killer cell-mediated cytotoxicity, and links between Pyk2 and Map Kinases) showed significant association with BCC risk with p value < 0.05 and false discovery rate (FDR) < 0.2. In sensitivity analyses, we changed the threshold of eSNP determination to 5×10^{-5} and 10^{-4} respectively, and the significant pathways identified in main analysis still ranked at top. Two positive controls in KEGG, the hedgehog signaling pathway and the BCC pathway, showed significant association with BCC risk in both main and sensitivity analyses. Our results show that SNPs that are undetectable by conventional GWASs are significantly associated with BCC when tested as pathways. Biological studies of these gene groups suggest their potential roles in the etiology of BCC.

Introduction

Basal cell carcinoma (BCC), a major type of non-melanoma skin cancer, is the most common malignancy among populations of European ancestry [1-3]. Though rarely fatal, the tumor may be locally invasive and cause clinically significant destruction of surrounding tissue if not treated adequately [4, 5]. In addition, subsequent skin cancers and other malignancies are more common among BCC patients in comparison to the general population [6].

Both environmental and genetic factors contribute to the genesis of BCC. Though exposure to ultraviolet (UV) radiation is generally accepted as the most important environmental risk factor for BCC, other known risk factors include family history of skin cancer and pigmentary characteristics, such as fair complexion, red or blond hair, and light eye color [7-9]. Most recently, genome-wide association studies (GWAS) have identified several genetic loci (including 1p36, 1q42, 5p15, 7q32, and 9q21, among others) associated with risk of BCC [10-12]. Despite the advances that have been made in understanding the etiology of BCC, the genetics of this complex disease is still largely unknown.

Although GWASs have revolutionized our ability to identify disease susceptibility loci or markers associated with them, they usually yield only the most significant SNPs, and the percentage of genetic variation explained by GWAS signals has generally been modest [13, 14]. One of the potential explanations for this "missing" heritability is that most common DNA variants with moderate effect size have not yet been identified by GWAS because of a lack of power [15]. Given this limitation of conventional association analysis, new approaches are

emerging to enhance the information extracted from current GWAS data. Pathway analysis, which jointly considers multiple variants with moderate signals in related genes, is a good complement to single-locus GWAS [16]. There is growing evidence that complex molecular networks and cellular pathways are often involved in disease susceptibility and disease progression [17, 18]. Thus, by taking into account prior biological knowledge about genes and pathways, we may have a better chance to identify disease-relevant loci [19], even though the signals individually do not meet the GWAS significance threshold [16].

Borrowing ideas from gene set enrichment analysis (GSEA) in the gene expression microarray field [20], Wang *et al.* first proposed pathway-based analysis of GWAS data in 2007 [16]. They used SNPs that are physically located in the gene region as the representative SNPs for that particular gene. However, SNPs within a gene region may not be the functional variants of the gene, and a gene may be regulated *in trans* by genetic variants that are physically distant [21]. Having realized this major shortcoming of conventional pathway analysis, as well as the importance of genetic variants that regulate gene transcription in mapping human disease genes [22], Zhong *et al.* suggested integration of expression-related SNPs (eSNPs) into conventional pathway analysis [23]. Two main aspects of this new approach are appealing: first, it further improves the power to detect genetic associations, because eSNPs can be considered functionally relevant variants [24]; secondly, it improves the interpretation of results, because variants that cluster within common biological pathways are taken into account jointly. This method has recently shown its potential strength in the context of type 2 diabetes GWAS [25]; however, applications to cancer have rarely been reported.

In 2012, Zhang *et al*. applied this novel pathway analysis to the GWAS of basal cell carcinoma for the first time [26]. Though that study provided novel insights into the biology underlying BCC, the false discovery rates of the identified pathways are of only marginal significance. Moreover, they used eSNPs discovered in two GWASs of global gene expression in lymphoblastoid cell lines (LCL) [22], which is not a tissue relevant to BCC. Because tissue dependency seems to be an important feature of disease susceptibility variants that regulate gene expression [27], ideally skin eSNPs should be used in BCC studies. Recently, the Multiple Tissue Human Expression Resource (MuTHER) project published detailed genomic and transcriptome data on three disease-relevant tissues (adipose, LCLs, and skin) originating from a cohort of 856 deeply phenotyped twins [28]. In the current study, we conducted a skin eSNPs-integrated pathway analysis for GWAS on BCC using MuTHER resources and sought to provide more insights into the underlying mechanisms of BCC.

Methods

Study populations

A BCC GWAS has been established within the sub-cohort of participants who provided a blood sample in Harvard cohorts. Eight case-control studies nested within the Nurses' Health Study (**NHS**), the Nurses' Health Study II (**NHS2**), and the Health Professionals Follow-up Study (**HPFS**) were included in the current BCC GWAS: the postmenopausal invasive breast cancer case-control study nested within the NHS (**BC_NHS**), the type 2 diabetes case-control studies nested within the NHS and the HPFS (**T2D_NHS** & **T2D_HPFS**), the coronary heart

disease case-control studies nested within the NHS and the HPFS (**CHD_NHS** & **CHD_HPFS**), and the kidney stone case-control studies nested within the NHS, the NHS2, and the HPFS (**KS_NHS, KS_NHS2** & **KS_HPFS**). See ***supplementary materials*** for more detailed descriptions of NHS, NHS2, HPFS, and the eight nested case-control studies. The study protocol was approved by the Institutional Review Boards of Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health.

Inclusion and exclusion

BCC cases who had other common cancers before diagnosis of BCC were excluded. Eligible controls were free of BCC and other common cancers. According to the National Cancer Institute and the American Cancer Society, common cancers include melanoma, SCC, breast cancer, endometrial cancer, ovarian cancer, colorectal cancer, bladder cancer, lung cancer, pancreatic cancer, kidney (renal cell) carcinoma, leukemia, non-Hodgkin lymphoma, thyroid cancer, and oral cancer. Participants with identical genetic information but different cohort IDs were removed; participants whose data appeared in more than one of the eight case-control studies were included only once. In total, the BCC GWAS comprised 2,323 BCC cases and 7,275 controls of European ancestry in the United States.

Genotyping, quality control (QC), and imputation

Samples from BC_NHS were genotyped using Illumina HumanHap550 array as part of the National Cancer Institute's Cancer Genetic Markers of Susceptibility (CGEMS) Project [29]. We

used Affymetrix 6.0 arrays for the T2D_NHS, T2D_HPFS, CHD_NHS, and CHD_HPFS, and Illumina 610Q for the KS_NHS, KS_NHS2, and KS_HPFS. Quality control on SNP completion rate, sample completion rate, deviation from Hardy-Weinberg equilibrium (HWE), Mendelian consistency, minor allele frequency, and duplication samples were conducted within each study, although the thresholds were chosen slightly differently. Within each of the eight studies, we used the MACH program [30] to impute genotypes for more than 2.5 million markers, using haplotype information in the HapMap phase II data build 36(CEU) as a reference panel.

BCC ascertainment

Disease follow-up procedures are identical for NHS, NHS2, and HPFS. Self-reported BCC case-control status is updated every two years without further pathological confirmation. The latest update was made in 2008 for the current analysis.

Multiple Tissue Human Expression Resource (MuTHER) project and eSNPs in skin tissue

A detailed description has been published previously [28]. Briefly, the MuTHER project included 856 female individuals of European ancestry recruited from the TwinsUK Adult twin registry [31]. Skin tissues were obtained from a photo-protected area adjacent to the umbilicus by punch biopsies. RNA from skin samples was extracted using TRIzol Reagent (Invitrogen), followed by RNA quality assessment and concentration measurement. Illumina Human Ht-12 V3 BeadChip (48,804 probes) was used for expression profiling of each sample, with either two or

three technical replicates. After quality control, expression profiling of skin tissue was performed on 705 individuals, and 23,596 probes were kept for further analysis. The TwinsUK study was genotyped by a combination of Illumina HumanHap300, HumanHap610Q, 1M-Suo, and 1.2M Duo 1M chips. Genetic imputation was carried out using IMPUTE software package and two reference panels: P0 [HapMap 2, release 22, combined Utah residents of Northern and Western European ancestry (CEU), Yoruba from Ibadan, Nigeria (YRI) and Asian (ASN) panels] and P1 (610k+, including the combined HumanHap610k and 1M arrays). Association of expression levels with probabilities of imputed genotypes were tested using a two-step mixed model-based score test [32, 33] and implemented in the GenABEL/ProbABEL package [34, 35] for 2,029,988 SNPs with MAF of >5% and IMPUTE info value of >0.8. In total, 667 skin samples that had both expression profiles and imputed genotypes were included in the analysis. Results of testing associations between gene expression level and SNPs were published and made publicly accessible on MuTHER's website in 2012 (http://www.muther.ac.uk/Data.html). We used a significance level of 10^{-5} for eSNP selection.

Statistical analysis

Association analysis: We used a multivariate logistic regression model, adjusted for age and the first three principal components of genetic variation, to evaluate the associations between eSNPs and BCC risk in each of the eight nested case-control studies. The principal components were calculated for all individuals on the basis of ca. 10,000 unlinked markers using the EIGENSTRAT software [36]. The within-study association results for each of the eSNPs were

combined by implementing inverse variance-weighted meta-analyses in METAL software [37].

eSNP enrichment analysis: We integrated the eSNP information into pathway-based GWAS analysis using the method of Zhong *et al.* [23]. For a gene whose expression is associated with multiple eSNPs, we chose the eSNP that had the most significant association with BCC risk as this gene's representative. Then we assigned these genes into the pathway defined by pathway databases. We evaluated the association of each pathway with risk of BCC with an Enrichment Score (ES), which was calculated from the weighted Kolmogorov-Smirnov-like running-sum statistic. This ES reflects the overrepresentation of genes within this pathway at the top of the entire ranked list of genes being tested. We permuted the case-control status and re-calculated the statistic values 1,000 times to assess the significance of each ES. To allow direct comparison of pathways of different sizes, a normalized enrichment score (NHS) was computed for each pathway. The FDR was calculated to estimate the proportion of false positive findings by using NES [38]. We set the significance level for the pathway analysis as p-value < 0.05 and FDR < 0.2.

Pathway databases: We used human biological pathways as defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/pathway.html/) database [39] as the primary pathway collection. Gene Ontology (GO, http://geneontology.org/) and BioCarta (http://www.biocarta.com/) databases were also included as secondary pathway collections. All pathways that contain at least 3 but at most 200 genes represented by eSNPs were tested.

Sensitivity analysis

Results (p-values) of all tested SNP-gene expression pairs are published on the MuTHER website. The threshold to identify SNPs that are significantly associated with at least one gene's expression in skin tissue is arbitrary. As the threshold becomes less stringent, the number of genes that can be represented by eSNPs increases and the surrogate eSNP for a particular gene may change. Therefore, we changed our threshold for eSNP selection to 5×10^{-5} and 10^{-4} respectively for the purpose of sensitivity analysis.

Results

From the MuTHER data, we identified 70,932, 87,481, and 97,903 eSNPs in skin tissue using the threshold of 10^{-5} (main analysis), 5×10^{-5} (sensitivity 1), and 10^{-4} (sensitivity 2) respectively. Among them, 69,988, 86,325, and 96,603 are available in our BCC GWAS, respectively. Because all these eSNPs have MAF >1% and imputation R-square >0.4 in our BCC GWAS, they were used for further analysis.

In our main analysis, 2,049 genes with surrogate eSNPs were assigned to the pathways defined in the KEGG database. Using the cut-off of containing 3-200 genes, 143 pathways were tested for association with BCC risk using our GWAS data. Eleven pathways reached a nominal p value < 0.05, which was 1.54-fold higher than the number expected by chance ($0.05\times143 = 7.15$; this is a conservative estimate, because pathways may be correlated due to overlapping genes, and the effective number should be smaller than 143). Three out of the 11 pathways had a FDR <0.2: the colorectal cancer pathway (p-value<0.00001, FDR =0.005), the regulation of actin

cytoskeleton pathway (p-value=0.03, FDR =0.073), and the basal cell carcinoma pathway (p-value=0.002, FDR =0.069). In sensitivity 1 analysis, the numbers of genes that can be represented by eSNPs increased to 2,649 when we used the threshold of 5×10^{-5} for eSNP identification. A total of 151 KEGG pathways that contain between 3 and 200 genes were examined for their association with BCC. Twelve reached a nominal $p<0.05$, which was 1.59-fold higher than the number expected by chance. Five out of the 12 pathways had a FDR <0.2. Besides the three that have already been found in the main analysis, the other two pathways are the adherens junction pathway (p-value=0.028, FDR =0.145) and the pancreatic cancer pathway (p-value=0.023, FDR =0.189). In sensitivity 2 analysis, 3,158 genes were included, and 164 KEGG pathways were tested. Fifteen reached a nominal $p<0.05$, which was 1.83-fold higher than the number expected by chance. Only one out of the 15 pathways had a FDR < 0.2 -- the colorectal cancer pathway (p-value<0.00001, FDR =0.175). In total, five KEGG pathways have shown significant associations with risk of BCC in either main analysis or sensitivity analysis. Results of main and sensitivity analyses for the five significant pathways are listed in ***Table 1.1.*** We also used GO and BioCarta databases for pathway construction. The results are shown in ***Tables 1.2 and 1.3***.

For certain pre-defined pathways identified through the pathway databases, only some of the genes could be represented by eSNPs. Therefore, more attention should be given to the genes and eSNPs that were included in the gene set enrichment analysis rather than to the entire pathway. For significant pathways, we summarized information on such genes and their corresponding eSNPs in ***Table 1.4***. Because no BioCarta pathway appeared to be significantly associated with

Table 1.1 KEGG Pathways with significant enrichment (p<0.05, FDR <0.2) in BCC GWAS & Hedgehog Signaling Pathway

Pathway	Gene count [d]	Main analysis [a]			Sensitivity analysis 1 [b]			Sensitivity analysis 2 [c]		
		Size [%]	Pathway enrichment p-value [e]	FDR[f]	Size [%]	Pathway enrichment p-value [e]	FDR[f]	Size [%]	Pathway enrichment p-value [e]	FDR[f]
Colorectal Cancer	114	7	***<0.00001***	***0.005***	10	***0.003***	***0.172***	12	***<0.00001***	***0.175***
Regulation of Actin Cytoskeleton	276	14	***0.03***	***0.073***	18	***0.03***	***0.183***	27	0.529	0.952
Basal Cell Carcinoma	73	3	***0.002***	***0.069***	4	***0.001***	***0.169***	4	<0.00001	0.269
Adherens Junction	110	7	0.346	1	10	***0.028***	***0.145***	11	0.02	0.253
Pancreatic Cancer	115	3	0.054	0.163	5	***0.023***	***0.189***	7	<0.00001	0.213
Hedgehog Signaling Pathway	74	3	0.008	0.657	5	0.031	0.464	5	0.036	0.404

a eSNPs were selected at significance level of 10^{-5}

b eSNPs were selected at significance level of 5×10^{-5}

c eSNPs were selected at significance level of 10^{-4}

d The number of genes in the pathway according to the KEGG database

e&f Based on 1,000 permutations

f Based on 143, 151, and 164 pathways in main, sensitivity 1, and sensitivity 2, respectively.

% The number of genes that have surrogate eSNPs in the pathway.

Table 1.2 GO Pathways with significant enrichment (p<0.05, FDR <0.2) in BCC GWAS

Pathway [#]	Gene count [d]	Main analysis [a]			Sensitivity analysis 1 [b]			Sensitivity analysis 2 [c]		
		Size [%]	Pathway enrichment p-value [e]	FDR[f]	Size [%]	Pathway enrichment p-value [e]	FDR[f]	Size [%]	Pathway enrichment p-value [e]	FDR[f]
GO0006921	42	3	***0.007***	***0.179***	4	0.042	0.932	5	0.137	0.941
GO0006997	70	4	***0.025***	***0.120***	5	0.099	0.717	7	0.166	0.902

a eSNPs were selected at significance level of 10^{-5}

b eSNPs were selected at significance level of 5×10^{-5}

c eSNPs were selected at significance level of 10^{-4}

d The number of genes in the pathway according to the GO database

e&f Based on 1,000 permutations

f Based on 407, 456, and 506 pathways in main, sensitivity 1, and sensitivity 2, respectively.

Annotation: GO0006921 – cellular component disassembly involved in apoptosis; GO0006997 – nucleus organization: a process at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of the nucleus

% The number of genes that have surrogate eSNPs in the pathway.

Table 1.3 BioCarta Pathways with significant enrichment (p<0.05, FDR <0.2) in BCC GWAS

Pathway [#]	Gene count [d]	Main analysis [a]			Sensitivity analysis 1 [b]			Sensitivity analysis 2 [c]		
		size	Pathway enrichment p-value	FDR	size	Pathway enrichment p-value	FDR	Size [%]	Pathway enrichment p-value [e]	FDR[f]
rasPathway	23	NA [+]			NA [+]			3	***0.008***	***0.109***
tcrPathway	45	NA [+]			NA [+]			6	***0.014***	***0.189***
nkcellsPathway	20	NA [+]			NA [+]			4	***0.024***	***0.188***
Pyk2Pathway	27	NA [+]			NA [+]			4	***0.048***	***0.199***

a eSNPs were selected at significance level of 10^{-5}

b eSNPs were selected at significance level of 5×10^{-5}

c eSNPs were selected at significance level of 10^{-4}

d The number of genes in the pathway according to the BioCarta database

e&f Based on 1,000 permutations

f Based on 60, 71, and 114 pathways in main, sensitivity 1, and sensitivity 2, respectively.

\+ These four pathways were not tested in main and sensitivity 1 analyses because their sizes are not between 3 and 200.

\# Annotation: rasPathway – Ras signaling pathway; tcrPathway – T cell Receptor signaling pathway; nkcellsPathway -- Ras-Independent pathway in NK cell-mediated cytotoxicity; Pyk2Pathway -- Links between Pyk2 and Map Kinases

% The number of genes that have surrogate eSNPs in the pathway.

Table 1.4 Genes and eSNPs in significant pathways identified in main analysis [&]

Pathway database	Pathway	Number of Genes with eSNP	Pathway enrichment *p-value*	FDR	Genes with eSNP	Chr [&&]	Surrogate eSNP [+]	eSNP P_{BCC} [#]	Chr_position [##]
KEGG	Colorectal Cancer	7	<0.00001	0.005	BIRC5	17	rs4789559	0.130	17:76218857
					CYCS	7	rs39454	0.025	7:25135783
					FZD3	8	rs12678890	0.075	8:28451002
					FZD8	10	rs11815242	0.101	10:35995340
					MAPK9	5	rs3812067	0.104	5:179709154
					SMAD3	15	rs7176870	0.097	15:67388553
					SOS1	2	rs12473092	0.029	2:39204040
	Regulation of Actin Cytoskeleton	14	0.03	0.073	ACTG1	17	rs12952655	0.717	17:80421139
					ARHGEF7	13	rs7984371	0.039	13:111958666
					BAIAP2	17	rs4969387	0.309	17:79081724
					C3orf10	3	rs279545	0. 051	3:9972493
					CYFIP2	5	rs11744003	0.085	5:156806993
					FGFR4	5	rs422421	0.099	5:176517326
					GNA12	7	rs7790322	0.051	7:2830498
					ITGA2	5	rs3212544	0.040	5:52358887
					ITGAX	16	rs11150612	0.103	16:31357760
					MYL2	12	rs16941319	0.593	12:111646853
					PAK2	3	rs7646247	0.431	3:196519209
					SOS1	2	rs12473092	0.029	2:39204040
					TIAM1	21	rs2833271	0.280	21:32487749
					VAV3	1	rs11185131	0.604	1:108078183
					VCL	10	rs12360087	0.002	10:76373904

Table 1.4 Genes and eSNPs in significant pathways identified in main analysis [&] (Continued)

Pathway database	Pathway	Number of Genes with eSNP	Pathway enrichment *p-value*	FDR	Genes with eSNP	Chr [&&]	Surrogate eSNP [+]	eSNP P_{BCC} [#]	Chr_position [##]
KEGG	Basal Cell Carcinoma	3	0.002	0.069	BMP2	20	rs6054443	0.001	20:6647580
					FZD3	8	rs12678890	0.075	8:28451002
					FZD8	10	rs11010260	0.051	10:35995340
GO	GO0006921	3	0.007	0.179	BIRC7	20	rs1075557	0.014	20:61870465
					CYCS	7	rs39454	0.025	7:25135783
					DFFB	1	rs4074709	0.802	1:3796948
	GO0006997	4	0.025	0.120	BIRC7	20	rs1075557	0.014	20:61870465
					CYCS	7	rs39454	0.025	7:25135783
					DFFB	1	rs4074709	0.802	1:3796948
					PML	15	rs11072463	0.199	15:74303349
BioCarta	rasPathway	3	0.008	0.109	PIK3R1	5	rs9291926	0.016	5:67599656
					RAC1	7	rs2689420	0.013	7:6410321
					RALGDS	9	rs482670	0.362	9:136007358
	tcrPathway	6	0.014	0.189	CALM3	19	rs973679	0.401	19:47061564
					NFATC2	20	rs231583	0.490	20:49346881
					NFATC3	16	rs13338993	0.289	16:67515312
					PIK3R1	5	rs9291926	0.016	5:67599656
					RAC1	7	rs2689420	0.013	7:6410321
					SOS1	2	rs12473092	0.029	2:39204040

Table 1.4 Genes and eSNPs in significant pathways identified in main analysis [&] (Continued)

Pathway database	Pathway	Number of Genes with eSNP	Pathway enrichment *p-value*	FDR	Genes with eSNP	Chr [&&]	Surrogate eSNP [+]	eSNP P_{BCC} [#]	Chr_position [##]
BioCarta	nkcellsPathway	4	0.024	0.188	PIK3R1	5	rs9291926	0.016	5:67599656
					PTK2B	8	rs472865	0.882	8:26698471
					RAC1	7	rs2689420	0.013	7:6410321
					SYK	9	rs914925	0.766	9:93584793
	Pyk2Pathway nkcellsPathway	4	0.048	0.199	CALM3	19	rs973679	0.401	19:47061564
					PTK2B	8	rs472865	0.882	8:26698471
					RAC1	7	rs2689420	0.013	7:6410321
					SOS1	2	rs12473092	0.029	2:39204040

& For the BioCarta database, results of sensitivity analysis 2 are presented in this table, because no significant pathway has been identified in main and sensitivity 1 analysis.

&& Chromosome of genes

\+ If a gene's expression is associated with multiple eSNPs, we used the eSNP that was most significantly associated with BCC risk as the gene's surrogate eSNP.

\# P_{BCC} represents P values of the association between eSNPs and risk of BCC.

Chromosome and position of eSNPs

BCC risk in main analysis (***Table 1.3***), we reported the results of sensitivity analysis 2 for BioCarta in ***Table 1.4***. On the other hand, some genes belong to more than one of the significant pathways. For example, SOS1 and RAC1 were included in four significant pathways and PIK3R1 and CYCS were in three significant pathways. Nine eSNPs associated with BCC risk at a nominal $P < 0.05$ are worth noting. These gene-eSNP pairs are CYCS-rs39454 ($P_{BCC} = 0.025$), SOS1-rs12473092 ($P_{BCC} = 0.029$), ARHGEF7- rs7984371 ($P_{BCC} = 0.039$), ITGA2- rs3212544 ($P_{BCC} = 0.040$), VCL – rs12360087 ($P_{BCC} = 0.002$), BMP2- rs6054443 ($P_{BCC} = 0.001$), BIRC7-rs1075557 ($P_{BCC} = 0.014$), PIK3R1- rs9291926 ($P_{BCC} = 0.016$), and RAC1-rs2689420 ($P_{BCC} = 0.013$).

Moreover, we chose two established BCC-related pathways in the KEGG database as positive controls – the basal cell carcinoma pathway and the hedgehog signaling pathway. The gene set enrichment p-value for these two pathways reached nominal significance in both the main and sensitivity analyses, though the FDRs of the hedgehog signaling pathway are above 0.2 (***Table 1.1***).

Discussion

Conventional GWASs have primarily focused on the associations between individual genetic markers and risk of diseases. In the current study, we applied a novel approach that integrates skin eSNPs and pathway analysis for GWAS of BCC. Three KEGG pathways (colorectal cancer, regulation of actin cytoskeleton, and basal cell carcinoma), two GO pathways (cellular component disassembly involved in apoptosis, and nucleus organization), and four BioCarta

pathways (Ras signaling pathway, T-cell receptor signaling pathway, Ras-independent pathway in natural killer (NK) cell-mediated cytotoxicity, and links between Pyk2 and Map Kinases) showed significant associations with BCC risk. Our results demonstrate that SNPs and genes of moderate effect that are undetectable by conventional GWASs are significantly associated with risk of BCC as groups. These gene sets might be implicated in the etiology of BCC.

Some well-known cancer-related pathways have been mapped in both the colorectal cancer pathway and the BCC pathway in KEGG, including the p53 signaling pathway, the Wnt signaling pathway, the PI3K-Akt signaling pathway, the TGF-β signaling pathway, and other pathways related to cell cycle and survival. Studies have shown that a personal history of non-melanoma skin cancer was significantly associated with a higher risk of other primary cancers [6, 40]. Certain genetic components may act systemically and play a role in both cutaneous and internal carcinogenesis. The actin cytoskeleton pathway mainly regulates cell motility, which is required for many biological processes, such as embryonic morphogenesis, immune surveillance, and tissue repair and regeneration. Aberrant regulation of cell migration drives progression of many diseases, including cancer invasion and metastasis [41, 42]. In the GO database, GO0006921 is defined as the breakdown of structures such as organelles, proteins, or other macromolecular structures during apoptosis; GO0006997 is defined as a process that is carried out at the cellular level that results in the assembly, arrangement of constituent parts, or disassembly of the nucleus, all of which are highly related to cancer development. The RAS signaling pathway is a key regulator of normal cell growth and malignant transformation. Mutations in RAS genes or alterations in upstream or downstream signaling components have

been found in most human tumors [43] including basal cell carcinoma, although with a relatively low mutation rate [44]. T cell receptor (TCR) activation promotes a number of signaling cascades that ultimately determine cell survival, proliferation, and differentiation. High levels of intratumor infiltration of T cells is correlated with prolonged survival in cancer patients [45]. NK cells are large granular lymphocytes with natural cytotoxicity against tumor cells [46]. An 11-year follow-up study has shown that low NK cell activity in peripheral blood is associated with increased cancer risk [47].

In the current study, we made a major improvement by using high-quality eSNPs data on disease-relevant tissue. Although detailed gene-expression studies have profiled transcripts and genotyped SNPs across the human genome in several population-based cohorts, gene expression data in skin tissue from a fairly large cohort was not accessible until the publication of the MuTHER project. In that study, the GWAS data and expression data had undergone stringent quality controls before testing the association of expression levels with probabilities of imputed genotypes. Also, skin eSNP identified in the MuTHER study had been replicated in independent cohorts [12]. Other strengths of our study include involvement of multiple pathway databases and design of sensitivity analysis as well as positive controls to validate our findings.

The main limitation of our study is that the proportion of genes that could be represented by eSNPs within a predefined pathway is too small, because only 69,988 SNPs that were significantly associated with expression of 2,049 genes at significance level of 10^{-5} had been included in the main analysis. For example, the colorectal cancer pathway in the KEGG database is composed of 114 genes, whereas only 7 genes (6%) were involved in the gene set enrichment

analysis. Specifically, we found that a subgroup of seven genes – BIRC5, CYCS, FZD3, FZD8, MAPK9, SMAD3, and SOS1 – that belong to the KEGG colorectal cancer pathway showed significant association with risk of BCC. Similar conclusions could be drawn for other significant pathways, with the subgroups presented in ***Table 1.4***. Given that the identified subgroups could hardly represent the original KEGG, GO, and BioCarta pathways, some may argue the necessity of using these pathway resources. However, these pre-defined pathways are important in two ways: on the one hand, they provide us prior knowledge on how to assign genes into different groups in order to conduct a pathway-based analysis; on the other hand, genes have been carefully selected, organized, and mapped in these established pathways based on multiple sources of evidence. With high-quality pre-collected information, we could interpret a gene's role and its relationship with other genes in the same pathway more easily, despite the limited size of identified subgroups.

A further limitation is that no replication was conducted for the identified gene groups, because we used all our BCC GWAS at the discovery stage to maximize statistical power. However, the significant gene groups in the main analysis also ranked top among all pathways being tested in sensitivity analyses. Besides, the positive controls – the Hedgehog signaling pathway and the BCC pathway – were significantly associated with risk of BCC (p<0.05) in both main and sensitivity analyses.

In conclusion, our study identified novel genes and gene sets that may be important for BCC development. Genes with moderate effect that are undetectable in conventional GWAS were significantly associated with risk of BCC as groups. Further pathway analyses that integrate

more skin eSNPs and/or other functional variants are warranted to verify our findings, and additional biological studies are needed to better elucidate the roles of these genes and pathways in the etiology of BCC.

Acknowledgements

We are indebted to the participants in the NHS, NHS2, and HPFS for their dedication to this research. We thank the following state cancer registries for their help: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington, and Wyoming. The authors assume full responsibility for analyses and interpretation of these data. This work is supported by NIH R01 CA49449, P01 CA87969, UM1 CA186107, and UM1 CA167552.

References

1. Wong, C., R. Strange, and J. Lear, *Basal cell carcinoma.* BMJ: British Medical Journal, 2003. **327**(7418): p. 794.
2. Diepgen, T. and V. Mahler, *The epidemiology of skin cancer.* British Journal of Dermatology, 2002. **146**(s61): p. 1-6.
3. Miller, D.L. and M.A. Weinstock, *Nonmelanoma skin cancer in the United States: incidence.* Journal of the American Academy of Dermatology, 1994. **30**(5): p. 774-778.
4. Chinem, V.P. and H.A. Miot, *Epidemiology of basal cell carcinoma.* Anais Brasileiros de Dermatologia, 2011. **86**(2): p. 292-305.
5. Zhao, B. and Y.-Y. He, *Recent advances in the prevention and treatment of skin cancer using photodynamic therapy.* Expert review of anticancer therapy, 2010. **10**(11): p. 1797-1809.
6. Song, F., et al., *Risk of a second primary cancer after non-melanoma skin cancer in white men and women: a prospective cohort study.* PLoS medicine, 2013. **10**(4): p. e1001433.
7. Gallagher, R.P., et al., *Sunlight exposure, pigmentary factors, and risk of nonmelanocytic skin cancer: I. Basal cell carcinoma.* Archives of Dermatology, 1995. **131**(2): p. 157.
8. Lear, J., et al., *Risk factors for basal cell carcinoma in the UK: case-control study in 806 patients.* Journal of the Royal Society of Medicine, 1997. **90**(7): p. 371.
9. Han, J., G.A. Colditz, and D.J. Hunter, *Risk factors for skin cancers: a nested case–control study within the Nurses' Health Study.* International journal of epidemiology, 2006. **35**(6): p. 1514-1521.

10. Stacey, S.N., et al., *Common variants on 1p36 and 1q42 are associated with cutaneous basal cell carcinoma but not with melanoma or pigmentation traits.* Nature genetics, 2008. **40**(11): p. 1313-1318.

11. Stacey, S.N., et al., *New common variants affecting susceptibility to basal cell carcinoma.* Nature genetics, 2009. **41**(8): p. 909-914.

12. Nan, H., et al., *Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma.* Human molecular genetics, 2011. **20**(18): p. 3718-3724.

13. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: potential next steps on a genetic journey.* Human molecular genetics, 2008. **17**(R2): p. R156-R165.

14. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

15. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* science, 2008. **322**(5903): p. 881-888.

16. Wang, K., M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* The American Journal of Human Genetics, 2007. **81**(6): p. 1278-1283.

17. Cordell, H.J., *Detecting gene–gene interactions that underlie human diseases.* Nature Reviews Genetics, 2009. **10**(6): p. 392-404.

18. Schadt, E.E., *Molecular networks as sensors and drivers of common human diseases.* Nature, 2009. **461**(7261): p. 218-223.

19. Wang, K., M. Li, and H. Hakonarson, *Analysing biological pathways in genome-wide association studies.* Nature Reviews Genetics, 2010. **11**(12): p. 843-854.

20. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

21. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS biology, 2008. **6**(5): p. e107.

22. Dixon, A.L., et al., *A genome-wide association study of global gene expression.* Nature genetics, 2007. **39**(10): p. 1202-1207.

23. Zhong, H., et al., *Integrating pathway analysis and genetics of gene expression for genome-wide association studies.* The American Journal of Human Genetics, 2010. **86**(4): p. 581-591.

24. Li, L., et al., *Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma.* Frontiers in genetics, 2013. **4**.

25. Zhong, H., et al., *Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes.* PLoS genetics, 2010. **6**(5): p. e1000932.

26. Zhang, M., et al., *Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma.* Human genetics, 2012. **131**(4): p. 615-623.

27. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS genetics, 2011. **7**(2): p. e1002003.

28. Grundberg, E., et al., *Mapping cis-and trans-regulatory effects across multiple tissues in twins.* Nature genetics, 2012. **44**(10): p. 1084-1089.

29. Hunter, D.J., et al., *A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.* Nature genetics, 2007. **39**(7): p. 870-874.

30. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genetic epidemiology, 2010. **34**(8): p. 816-834.

31. Spector, T.D. and F.M. Williams, *The UK adult twin registry (TwinsUK).* Twin Research and Human Genetics, 2006. **9**(6): p. 899-906.

32. Aulchenko, Y.S., D.-J. de Koning, and C. Haley, *Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis.* Genetics, 2007. **177**(1): p. 577-585.

33. Chen, W.-M. and G.R. Abecasis, *Family-based association tests for genomewide association scans.* The American Journal of Human Genetics, 2007. **81**(5): p. 913-926.

34. Aulchenko, Y.S., et al., *GenABEL: an R library for genome-wide association analysis.* Bioinformatics, 2007. **23**(10): p. 1294-1296.

35. Aulchenko, Y.S., M.V. Struchalin, and C.M. van Duijn, *ProbABEL package for genome-wide association analysis of imputed data.* BMC bioinformatics, 2010. **11**(1): p. 134.

36. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.

37. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-2191.

38. Reiner, A., D. Yekutieli, and Y. Benjamini, *Identifying differentially expressed genes using false discovery rate controlling procedures.* Bioinformatics, 2003. **19**(3): p. 368-375.

39. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

40. Wheless, L., J. Black, and A.J. Alberg, *Nonmelanoma skin cancer and the risk of second primary cancers: a systematic review.* Cancer Epidemiology Biomarkers & Prevention, 2010. **19**(7): p. 1686-1695.

41. Sahai, E., *Mechanisms of cancer cell invasion.* Current opinion in genetics & development, 2005. **15**(1): p. 87-96.

42. Olson, M.F. and E. Sahai, *The actin cytoskeleton in cancer cell motility.* Clinical & experimental metastasis, 2009. **26**(4): p. 273-287.

43. Downward, J., *Targeting RAS signalling pathways in cancer therapy.* Nature Reviews Cancer, 2003. **3**(1): p. 11-22.

44. van der Schroeff, J.G., et al., *Ras oncogene mutations in basal cell carcinomas and squamous cell carcinomas of human skin.* Journal of investigative dermatology, 1990. **94**(4): p. 423-425.

45. Cronin, S.J. and J.M. Penninger, *From T‐cell activation signals to signaling control of anti‐cancer immunity.* Immunological reviews, 2007. **220**(1): p. 151-168.

46. Vivier, E., et al., *Functions of natural killer cells.* Nature immunology, 2008. **9**(5): p. 503-510.

47. Imai, K., et al., *Natural cytotoxic activity of peripheral-blood lymphocytes and cancer incidence: an 11-year follow-up study of a general population.* The Lancet, 2000. **356**(9244): p. 1795-1799.

Supplementary Materials

1. Detailed description of study population

Nurses' Health Study (NHS): The NHS is a prospective cohort study established in 1976 with 121,700 female U.S registered nurses, who were then 30-55 years old. All of them completed and returned a mailed self-administered questionnaire about their medical histories and lifestyle at the baseline. In 1989 and 1990, a total of 32,826 women provided blood samples. Information regarding medical history, lifestyle, and disease diagnoses was updated every 2 years with a follow-up rate of 90%.

Health Professionals Follow-up Study (HPFS): The HPFS began in 1986 with 51,529 U.S. male health professionals who were 40-75 years old at initial recruitment. They all answered a detailed mailed questionnaire at the inception of the study. Disease- and health-related information was obtained and updated through biennial questionnaires. Between 1993 and 1994, 18,159 of these men provided a blood sample. The average follow-up rate for this cohort over 10 years is greater than 90%.

Nurses' Health Study II (NHS2): The NHS2 was established in 1989, when 116,671 female registered nurses aged 25–42 and residing in the United States at the time of enrollment responded to an initial questionnaire on their medical histories and baseline health-related exposures. Information regarding medical history, lifestyle risk factors, and disease diagnoses was updated every 2 years with a follow-up rate of above 90%. Blood samples from 29,616 nurses were collected in the late 1990's.

BCC GWAS set: A BCC GWAS set has been established within the sub-cohort of participants who provided a blood sample. Eight case-control studies were included in current BCC GWAS, they are:

1) Postmenopausal invasive breast cancer case-control study nested within the NHS (**BC_NHS**): Eligible cases in this study consisted of women with pathologically confirmed incident breast cancer from the subcohort who gave a blood specimen. Cases with a diagnosis after blood collection up to June 1, 2000 with no previously diagnosed cancer except for non-melanoma skin cancer were included. One control for each case was randomly selected among women who gave a blood sample and were free of diagnosed cancer (excluding non-melanoma skin cancer) up to and including the interval in which the case was diagnosed. Controls were matched to cases on year of birth, menopausal status, recent post-menopausal hormone (PMH) use, month of blood return, time of day of blood collection, and fasting status at blood draw [1].

2) Type 2 diabetes (T2D) case-control study nested within the NHS and HPFS (**T2D_NHS** and **T2D_HPFS**): Diabetes cases were defined as self-reported incident diabetes confirmed by a validated supplementary questionnaire. For cases before 1998, diagnosis was made using criteria consistent with those proposed by the National Diabetes Data Group (NDDG). For cases during the 1998 and 2000 cycles, the American Diabetes Association's diagnostic criteria were used for the diagnosis of diabetes cases. The nondiabetic control subjects were matched to cases on age, month and year of blood draw, and fasting status [2].

3) Coronary heart disease (CHD) case-control study nested within the NHS and HPFS (**CHD_NHS** and **CHD_HPFS**): In both the NHS and HPFS, participants who had reported an

incident CHD event on the follow-up questionnaire were contacted for confirmation and permission to review medical records was requested. Medical records for deceased participants were also sought for deaths that were identified by families and postal officials and through the National Death Index. Physicians blinded to the participant's questionnaire reports reviewed all medical records. Fatal CHD cases were identified primarily through review of medical records [3]. Among participants who provided blood samples and who were without cardiovascular disease or cancer at blood draw, incident CHD cases occurring after blood draw were selected as cases. Controls were selected in a 2:1 ratio matched to cases on age, smoking, and month of blood return.

4) *Kidney stone case-control study nested within the NHS, NHS2 and HPFS (**KS_NHS, KS_NHS2** and **KS_HPFS**):* Participants from KS_NHS, KS-NHS2 and KS_HPFS were individuals who performed a 24-hour urine collection; two-thirds had a history of incident nephrolithiasis. Details regarding the urine collection [4] and the confirmation of kidney stone disease were published previously [5]. The biennial questionnaires have asked whether a participant had been diagnosed with kidney stone. For newly reported cases, an additional questionnaire was sent to inquire date of occurrence and symptoms. Studies have been conducted to confirm the validity of the self-reported stones [6, 7]. A control was randomly selected from the blood cohorts of NHS, NHS2, and HPFS for each case, matching on age, time of blood draw, and fasting status.

Supplementary Table 1.1 Number of BCC cases and controls [#] in the eight case-control studies nested in NHS, NHS2 or HPFS

Study	Number of BCC cases	Number of controls
BC_NHS	248	816
T2D_NHS	665	2162
T2D_HPFS	597	1555
CHD_NHS	253	765
CHD_HPFS	282	715
KS_NHS	99	324
KS_NHS2	58	552
KS_HPFS	121	386
Total	2323	7275

BCC cases who had diagnosis of other common cancers before diagnosis of BCC were excluded; Controls who had other cancers were excluded; participants with identical genetic information but different cohort ID were removed; participants sampled by more than one studies were included only once. Participants who withdrew consent were excluded.

References

1. Hunter, D.J., et al., *A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.* Nat Genet, 2007. **39**(7): p. 870-4.
2. Qi, L., et al., *Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes.* Hum Mol Genet, 2010. **19**(13): p. 2706-15.
3. Rimm, E.B., et al., *Prospective study of alcohol consumption and risk of coronary disease in men.* Lancet, 1991. **338**(8765): p. 464-8.
4. Curhan, G. and E. Taylor, *24-h uric acid excretion and the risk of kidney stones.* Kidney international, 2007. **73**(4): p. 489-496.
5. Taylor, E.N., M.J. Stampfer, and G.C. Curhan, *Obesity, weight gain, and the risk of*

kidney stones. Jama, 2005. **293**(4): p. 455-462.

6. Curhan, G.C., et al., *A prospective study of dietary calcium and other nutrients and the risk of symptomatic kidney stones.* New England Journal of Medicine, 1993. **328**(12): p. 833-838.

7. Curhan, G.C., et al., *Comparison of dietary calcium with supplemental calcium and other nutrients as factors affecting the risk for kidney stones in women.* Annals of Internal Medicine, 1997. **126**(7): p. 497-504.

CHAPTER 2

A genome-wide analysis of gene-caffeine consumption interaction on basal cell carcinoma

Xin Li [1], Marilyn C. Cornelis [2], Liming Liang [1], Immaculata De Vivo [1,3], Edward Giovannucci [1,3,4], Jean Y. Tang [5], and Jiali Han [6]

[1] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

[2] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA;

[3] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA;

[4] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

[5] Department of Dermatology, Stanford University School of Medicine, Redwood City, CA, USA;

[6] Department of Epidemiology, Fairbanks School of Public Health, Indiana University, and Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, IN, USA

Abstract

Increased caffeine consumption is significantly associated with reduced risk of basal cell carcinoma (BCC). To identify common genetic markers that may modify this association, we tested gene-caffeine intake interaction on BCC risk in a genome-wide analysis. We included 3,383 BCC cases and 8,528 controls of European ancestry in the Nurses' Health Study and Health Professionals Follow-up Study. We determined that SNP rs142310826 (minor allele frequency or MAF = 1.9%) shows a genome-wide significant interaction with caffeine consumption ($p = 1.78\times10^{-8}$ for interaction). The estimates for interaction between this SNP and caffeine intake in relation to BCC among women were not significantly different from that among men (p = 0.64 for heterogeneity). We also found several loci that modify the caffeine-BCC association differently in men and women (p for heterogeneity between genders < 0.001): genetic markers on chromosomes 2p12, 1q32.2, and 10p13 in women, and SNPs on chromosome 8p11.21 in men. A sensitivity analysis that modeled caffeine consumption in a different way did not change our results materially. This study is proof of concept that inclusion of environmental factors can help identify genes that are missed in conventional genome-wide association studies. Validation of these findings in additional populations may facilitate targeted BCC prevention strategies.

Introduction

Basal cell carcinoma (BCC), a major histological type of non-melanoma skin cancer, is the most common malignancy among populations of European ancestry [1, 2]. Some known risk factors for BCC include exposure to ultraviolet (UV) radiation, family history of skin cancer, and lighter pigmentation [3-5]. Like other common disorders, BCC is thought to have both environmental and genetic components and to involve their interactions. Genome-wide association studies (GWASs) have identified several genetic loci that confer susceptibility to BCC [6-8], however they tested only for association between individual genetic markers and risk of BCC without taking interactions into consideration.

Caffeine is the most widely consumed psychoactive substance in the world. Studies have demonstrated caffeine's protective role against the development of BCC. In mice, oral or topical administration of caffeine inhibits UV-induced carcinogenesis [9-11]. Consumption of coffee or tea has been associated with lower incidence of non-melanoma skin cancers in several epidemiological studies [12-14]. In a recent prospective study using data from the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS), researchers found a significant inverse association between total caffeine intake and risk of BCC [15]. However, the mechanisms behind this association are not well understood. One potential explanation is that caffeine could augment apoptosis in UV-damaged keratinocytes through ataxia telangiectasia and rad3-related (ATR) kinase and its downstream effector checkpoint kinase-1 (Chk1) [16-19]. However, conclusions from these biological studies may not readily be applied to the human

body, because they were based on mice and cultured cells to which a larger-than-usual dose of caffeine was administered.

To the best of our knowledge, no epidemiological study considering both caffeine consumption and the genetic makeup of participants has been conducted to investigate the interrelationship of caffeine and genetic markers in BCC development. In the current study, we conducted genome-wide analyses of gene-caffeine consumption interactions among the participants of the blood cohorts in the NHS and HPFS. Our study may provide new biologic insights into caffeine's role in BCC development, lead to discovery of BCC-related genes that have been missed in conventional GWASs, and help identify subgroups who might benefit from personalized advice concerning their coffee consumption habits.

Methods

Study population, inclusion, and exclusion

Eighteen case-control studies nested within the NHS and HPFS with cleaned genotype data were included in our study. Participants who had other common cancers before 1986 were excluded because we only considered caffeine intake measured in 1986. BCC cases who had other common cancers before diagnosis of BCC were excluded. Eligible controls were free of BCC and other common cancers. Participants with identical genetic information but different cohort IDs were removed. Participants who were sampled in more than one of the 18 case-control studies were included only once. In total, 3,383 BCC cases and 8,528 controls of

European ancestry were available. See ***Supplementary Materials*** for more detailed descriptions of the NHS, the HPFS, and the 18 nested case-control studies. The study protocol was approved by the Institutional Review Boards of Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health.

Genotyping, quality control, and imputation

Samples from the 18 nested case-control studies were genotyped using a variety of platforms. We combined these datasets into three compiled datasets based on their genotype platform type: Affymetrix (Affy), Illumina HumanHap series (Illumina), or Illumina Omni Express (Omni) (***Supplementary Table 2.1***). Quality control on SNP completion rate, sample completion rate, ancestry consistency, deviation from Hardy-Weinberg equilibrium (HWE), Mendelian consistency, minor allele frequency, and duplication samples were conducted within each of the three combined datasets. We then imputed the complied datasets using the 1000 Genomes Project ALL Phase I Integrated Release Version 3 Haplotypes excluding monomorphic and singleton sites (2010-11 data freeze, 2012-03-14 haplotypes) as a reference panel. Detailed descriptions of quality control and imputation are provided in ***Supplementary Materials***. We included genetic markers with imputation Rsq >=0.3 and minor allele frequency >= 1% in further analysis. The numbers of such markers in the three combined datasets are presented in ***Supplementary Table 2.2***.

Caffeine intake

Information on dietary intake of coffee and other foods known to be high in caffeine, including tea, cola, and chocolate, was collected by food-frequency questionnaire (FFQ). The questionnaires were completed in 1984, 1986, 1990, 1994, 1998, 2002, and 2006 for the NHS, and in 1986, 1990, 1994, 1998, 2002, and 2006 for the HPFS. On all questionnaires, participants were asked how many times on average during the previous year they had consumed each food and beverage. The participants could choose from nine frequency responses (never, 1-3 per month, 1 per week, 2-4 per week, 5-6 per week, 1 per day, 2-3 per day, 4-5 per day, and >=6 per day). Based on information obtained from the FFQ, the total intake of caffeine was calculated by multiplying the reported frequency of each food by the caffeine content of one serving of that food (1 cup for coffee or tea, one 12-ounce bottle or can for carbonated beverages, and 1 ounce for chocolate). According to the U.S. Department of Agriculture food composition sources, caffeine content is 137 mg per cup of caffeinated coffee, 47 mg per cup of tea, 46 mg per bottle or can of cola beverage, and 7 mg per serving of chocolate candy. Food and nutrient intakes assessed by this dietary questionnaire, including caffeine, have been validated previously against two 1-week diet records. The observed correlation between the questionnaire and the diet record was about 0.9 for caffeine consumption [20, 21]. We used daily caffeine intake (mg) measured in 1986 in the current study.

BCC ascertainment

Disease follow-up procedures are identical for the NHS and the HPFS. Self-reported BCC case-control status is updated every two years starting in 1984 in the NHS and 1986 in the HPFS

without further pathological confirmation. The latest updates were made in 2008 in the NHS and 2010 in the HPFS. The validity of self-reported BCC in these medically sophisticated populations has been assessed in previous studies [22].

Statistical analysis

To account for gender differences (cohort differences), we divided each of the combined datasets into two parts and conducted genome-wide gene-environment (G-E) interaction analysis in the six datasets (Illumina_NHS, Illumina_HPFS, Affy_NHS, Affy_HPFS, Omni_NHS, and Omni_HPFS). We used standard logistic regression with a product term to test the interaction between caffeine consumption and genetic markers in relation to BCC risk, adjusted for age in 1986, and the first three principal components (PCs) from EIGENSTRAT [23] to account for population substructure. Both genotyped and imputed markers were examined as continuous variables, assuming additive effects. Quartiles of caffeine intake were defined within each of the six datasets using the full range from zero to maximum intake among controls. We coded the quartiles as an ordinal variable (1^{st} quartile = 1, 2^{nd} quartile=2, 3^{rd} quartile =3, 4^{th} quartile = 4) in the main analysis, and used the median values of each quartile to represent the corresponding intake levels (1^{st} quartile = median intake of 1^{st} quartile, 2^{nd} quartile=median intake of 2^{nd} quartile, etc.) for sensitivity analysis. We combined results from Illumina_NHS, Affy_NHS, and Omni_NHS using inverse variance-weighted meta-analyses in METAL software [24]. The same procedure was implemented for the three HPFS datasets to obtain combined results for men. We calculated *p-value* for heterogeneity between men and women with the Cochran Q test, and

performed the third meta-analysis for all six datasets if no significant difference was found between genders. All analyses were conducted using the ProbABEL package [25] and R-3.0.2 (https://www.r-project.org) .

Results

The characteristics of participants within each of the six subsets are provided in ***Table 2.1***. Participants in the NHS (women) consumed more caffeine compared to those in the HPFS (men). BCC was more prevalent among males.

For our main analysis, in which quartiles of caffeine intake were modeled as ordinal variables, the *p-value* for interaction between each genetic marker and caffeine are shown in the Manhattan plot and quantile-quantile (Q-Q) plot (***Figure 2.1***). On chromosome 4, we determined that SNP rs142310826 (MAF = 1.9%) had a genome-wide significant interaction with caffeine consumption ($p = 1.78\times10^{-8}$ for interaction). The estimate for interaction between this SNP and caffeine intake in relation to BCC among women was not significantly different from that among men ($p = 0.64$ for heterogeneity). Using the UCSC GRCh37/hg19 assembly, this SNP was mapped to gene *NEIL3* (***Table 2.2***).

In the gender-specific analysis, 19 genetic markers and 3 genetic markers with *p-value* for interaction less than 5×10^{-7} were identified among women and men, respectively (***Table 2.3***). The top significant marker identified among females was 2:76738900:TTAGA ($p = 2.51\times10^{-8}$ for interaction), which was mapped on gene *LRRTM4*. Eleven genetic variants located very close to this top marker were also identified. As expected, the beta estimates, MAF, and imputation

quality of these related genetic markers are very similar. The other two regions identified in NHS were mapped to gene *ATF3* on chromosome 1 and gene *DCLRE1C* on chromosome 10, respectively. In men, three correlated SNPs on chromosome 8 (*POTEA* gene) were reported. However, the *p-value* for interaction of the most significant one, rs77868414, failed to reach genome-wide significance ($p = 5\times10^{-8}$). All 22 gender-specific markers reported above (19 in NHS and 3 in HPFS) had *p-values* for heterogeneity <0.001 or approximately 0.001. Therefore, we did not calculate the combined estimates for them. Manhattan plots and Q-Q plots for gender-specific analysis are shown in ***Supplementary Figures 2.1 and 2.2***. Markers that interact with caffeine consumption in relation to BCC risk at the significance level of 5×10^{-6} are presented in ***Supplementary Tables 2.3, 2.4, and 2.5***.

In sensitivity analysis, we used the median values of each quartile to represent the corresponding caffeine intake levels (1^{st} quartile = median intake of 1^{st} quartile, 2^{nd} quartile=median intake of 2^{nd} quartile, etc.), which did not change the results materially. ***Supplementary Table 2.6*** summarizes sensitivity analysis results for the independent markers discovered by main analysis.

Genome-wide association studies have identified several genetic loci that are associated with caffeine/coffee consumption [26-29]. We extracted results for these SNPs from our genome-wide G-E interaction analysis to better illustrate their potential functions. We also tested their individual and combined association with risk of BCC, but no significant association was found. ***Table 2.4*** shows analytic methods and results for these additional analyses.

Table 2.1 Descriptive Characteristics of Study Population

Datasets	No. (%)		Gender	Age in 1986, Mean [Min, Max]	Quartiles of caffeine intake, mg/day, Median [Min, Max] &			
	BCC cases	BCC controls			1st Quartile	2nd Quartile	3rd Quartile	4th Quartile
Illumina_NHS	544 (28.6%)	1355 (71.4%)	Female	54 [40, 65]	45 [0,93]	150 [94,208]	353 [209, 381]	554.5 [382, 1128]
Illumina_HPFS	302 (34.5%)	573 (65.5%)	Male	53 [40, 76]	14.5 [0,39]	81 [40,145]	201 [146, 358]	458 [359, 943]
Affy_NHS	785 (23.5%)	2556 (76.5%)	Female	54 [40, 65]	39 [0, 88]	160 [89, 229]	354 [231, 394]	630 [395, 1268]
Affy_HPFS	818 (31.9%)	1748 (68.1%)	Male	55 [40, 73]	11 [0, 48]	115 [49, 167]	348 [168, 377]	630 [378, 1114]
Omni_NHS	524 (25.7%)	1513 (74.3%)	Female	54 [40, 65]	42 [0, 85]	150 [88, 221]	356 [222, 408]	627 [409,1220]
Omni_HPFS	410 (34.4%)	783 (65.6%)	Male	54 [40, 75]	17 [0, 52]	120 [53, 166]	348 [167, 371]	542 [372, 1037]

& Quartiles of caffeine intake were defined for each dataset using the full range of intake (from zero to maximum) among BCC controls.

Table 2.2 Genetic markers with P-value for interaction < 5×10^{-8} in meta-analysis

SNP rs number	CHR:BP	A1	A2	Freq1[a]	Average Imputation Rsq [b]	NHS [c]		HPFS [d]		Meta-analysis [e]		P-Het [f]	Mapped Gene [g]
						Beta	P-value	Beta	P-value	Beta	P-value		
rs142310826	4:179402856	a	t	0.0185	0.580	-0.82	4.96E-06	-0.68	8.55E-04	-0.76	1.78E-08	6.39E-01	NEIL3

a Frequency for allele 1;
b Average imputation R square quality metric of Illumina, Affy, and Omni datasets;
c Results of meta-analysis of Illumina_NHS, Affy_NHS, and Omni_NHS;
d Results of meta-analysis of Illumina_HPFS, Affy_HPFS, and Omni_HPFS;
e Results of meta-analysis of Illumina_NHS, Affy_NHS, Omni_NHS, Illumina_HPFS, Affy_HPFS, and Omni_HPFS;
f P for heterogeneity between results of NHS and HPFS (gender difference);
g The SNP was mapped to its nearest genes using the UCSC GRCh37/hg19 assembly;

Table 2.3 Genetic markers with P-value for interaction < 5×10^{-7} in gender-specific analysis

SNP rs number	CHR:BP	A1	A2	Freq1 [a]	Average Imputati-on Rsq [b]	NHS [c]		HPFS [d]		P-Het [e]	Mapped Gene [f]
						Beta	P-value	Beta	P-value		
NHS (Female)											
NA [&]	2:76738900:TTAGA	d	r	0.1860	0.969	-0.27	2.51E-08	-0.02	6.77E-01	5.50E-04	LRRTM4
rs12624158	2:76736544	a	t	0.8186	0.984	0.26	9.62E-08	0.02	7.20E-01	8.42E-04	
rs12477078	2:76711828	t	c	0.1835	0.992	-0.25	1.18E-07	-0.003	9.47E-01	4.13E-04	
rs1921242	2:76704590	t	c	0.8146	0.995	0.25	1.53E-07	0.01	8.20E-01	6.97E-04	
rs12474352	2:76714536	c	g	0.8143	0.997	0.25	1.80E-07	0.01	8.15E-01	7.75E-04	
rs12466281	2:76714567	a	g	0.8143	0.997	0.25	1.81E-07	0.01	8.16E-01	7.79E-04	
rs12613882	2:76722019	a	g	0.1831	0.999	-0.25	1.93E-07	-0.003	9.60E-01	5.00E-04	
rs12618802	2:76725088	a	g	0.8141	0.998	0.24	2.60E-07	0.02	7.32E-01	1.24E-03	
rs17012789	2:76722305	t	c	0.8140	0.999	0.24	2.79E-07	0.01	7.82E-01	1.07E-03	
rs12476072	2:76702062	t	g	0.8152	0.992	0.24	3.78E-07	-0.004	9.44E-01	4.94E-04	
rs4853248	2:76692117	a	g	0.1828	0.987	-0.24	4.04E-07	0.01	8.22E-01	3.58E-04	
rs12463620	2:76701736	a	t	0.8155	0.993	0.24	4.16E-07	-0.001	9.78E-01	6.01E-04	
rs6694870	1:212728988	a	g	0.1528	0.957	0.27	1.64E-07	-0.03	5.53E-01	5.50E-05	ATF3
rs11119969	1:212729555	a	g	0.1527	0.957	0.27	1.67E-07	-0.03	5.47E-01	5.41E-05	
rs1344329	1:212724028	a	g	0.1442	0.962	0.28	1.74E-07	-0.03	6.27E-01	7.25E-05	
rs12073156	1:212721216	t	c	0.8556	0.968	-0.28	1.82E-07	0.02	6.70E-01	9.18E-05	
rs931449	1:212726219	a	g	0.1511	0.959	0.27	2.55E-07	-0.03	5.64E-01	7.40E-05	
rs12072138	1:212721010	t	c	0.8537	0.959	-0.27	2.62E-07	0.02	6.87E-01	1.24E-04	
rs191976747	10:15021717	a	t	0.0105	0.629	-1.32	4.51E-07	-0.04	8.80E-01	5.48E-04	DCLRE1C

Table 2.3 Genetic markers with P-value for interaction < 5×10^{-7} in gender-specific analysis (Continued)

SNP rs number	CHR:BP	A1	A2	Freq1 [a]	Average Imputati-on Rsq [b]	NHS [c]		HPFS [d]		P-Het [e]	Mapped Gene [f]
						Beta	P-value	Beta	P-value		
HPFS (Male)											
rs77868414	8:43285229	c	g	0.0343	0.377	0.06	7.19E-01	-0.97	3.43E-07	3.50E-05	POTEA
rs117285747	8:43290783	c	g	0.9625	0.411	-0.06	7.02E-01	0.88	4.01E-07	3.72E-05	
rs75594195	8:43273165	a	g	0.9635	0.416	-0.07	6.52E-01	0.88	4.72E-07	3.55E-05	

& This marker does not have a rs number

a Frequency for allele 1;

b Average imputation R square quality metric of Illumina, Affy, and Omni datasets;

c Results of meta-analysis of Illumina_NHS, Affy_NHS, and Omni_NHS;

d Results of meta-analysis of Illumina_HPFS, Affy_HPFS, and Omni_HPFS;

e P for heterogeneity between results of NHS and HPFS (gender difference);

f Genetic markers were mapped to their nearest genes using the UCSC GRCh37/hg19 assembly.

Table 2.4 Interaction between caffeine consumption-related SNPs and caffeine in relation to BCC risk; Individual and combined association between caffeine consumption-related SNPs and risk of BCC

Caffeine consumption-related SNPs [a]	CHR:BP	Mapped genes [b]	P_{BCC} &&	$P_{Interaction}$ ##	Functional group	Association between genetic scores by functional group and risk of BCC [c]	Association between the comprehensive genetic score and risk of BCC [c]
rs1481012	4:89039082	ABCG2	0.39	0.06	Caffeine metabolism	P %% = 0.50	P %% = 0.78
rs17685	7:75616105	POR	0.19	0.48			
rs2470893	15:75019449	CYP1A1	0.27	0.57			
rs2472297	15:75027880	CYP1A2	0.39	0.57			
rs6968554	7:17287106	AHR	0.97	0.16			
rs6968865	7:17287269	AHR	0.97	0.22			
rs6265	11:27679916	BDNF	0.01	0.18	Addiction	P %% = 0.32	
rs9902453	17:28349095	EFCAB5	0.43	0.25			
rs1260326	2:27730940	GCKR	0.55	0.08	unknown	NA ++	
rs7800944	7:73035857	MLXIPL	0.13	0.68			

a This SNP list was obtained from the GWAS catalog (http://www.ebi.ac.uk/gwas/home). For SNPs in strong LD (R^2>0.8), we only kept the one that is more significantly associated with caffeine intake in our dataset for following analysis;

b Reported by original GWAS papers;

c For each individual, we summed the dosage of alleles that associated with increased caffeine intake to obtain the genetic score. Three scores were calculated – the caffeine metabolism genetic score, the addiction genetic score, and the comprehensive genetic score;

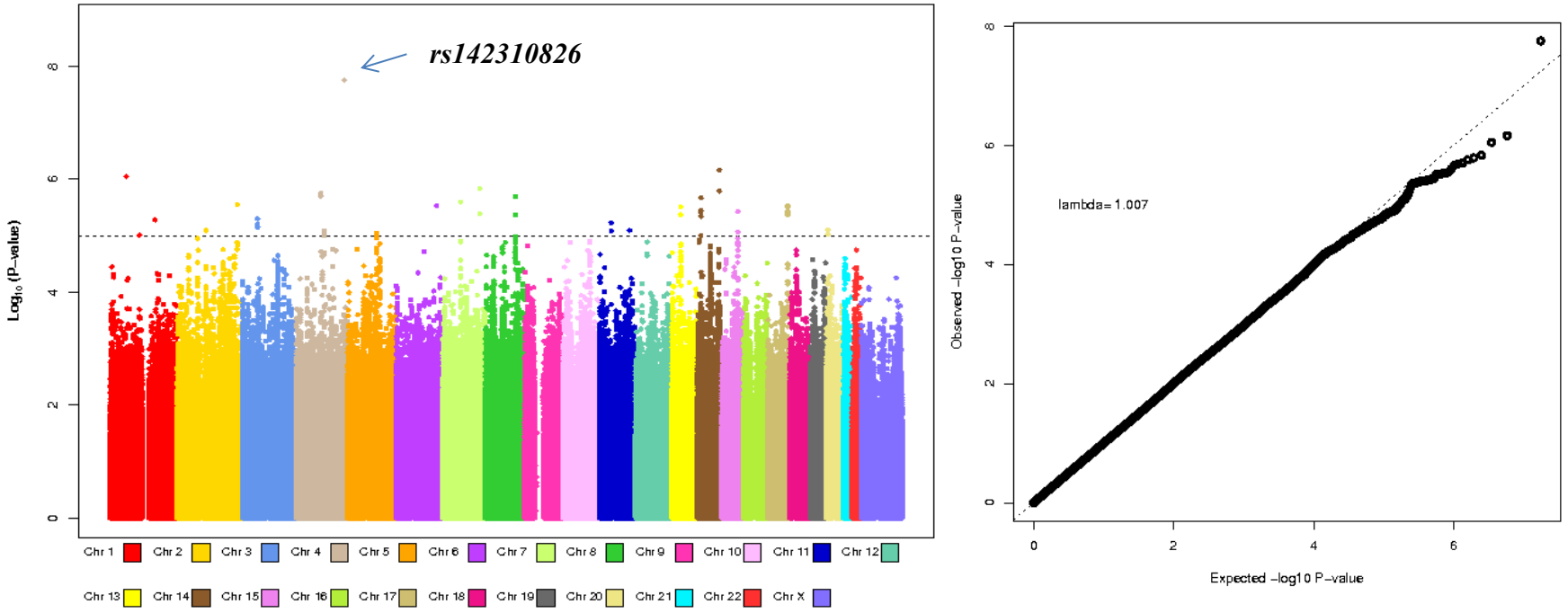
&& P-value for the association between SNPs and risk of BCC;

P-value for SNPs' interaction with caffeine intake in relation to risk of BCC;

%% P-value for the association between genetic scores and risk of BCC. Models adjusted for gender, sex, and top three PCs.

++ Functional genetic score was not calculated for SNPs of unknown function.

Figure 2.1 Manhattan plot and Q-Q plot for the interaction results with caffeine intake [&]



& Results are for the meta-analysis of six datasets. Quartiles of caffeine intake were modeled as an ordinal variable (main analysis).

Discussion

Using data from the NHS and the HPFS, our group has previously reported an inverse association between dietary caffeine intake and risk of BCC [15]. Compared with the lowest quintile, the highest quintile had the lowest risk (RR=0.82 in women, 95%CI: 0.76-0.86; RR=0.87 in men, 95% CI: 0.81-0.94; P_{trend} <0.0001 in both). In the current study, we conducted a meta-analysis of genome-wide G-E interaction studies among the blood cohort participants in the NHS and HPFS. We determined that caffeine consumption was differentially associated with BCC risk according to genetic variation at SNP rs142310826, which is located on chromosome 4q34.3. We also found several loci that modify the caffeine-BCC association differently in men and women (*p* for heterogeneity between genders < 0.001): genetic markers at chromosome 2p12, 1q32.2, and 10p13 in women, and SNPs at chromosome 8p11.21 in men. These genetic markers and their mapped genes may prove to be important in BCC etiology, especially when caffeine consumption is considered.

UV-induced DNA damage in skin cancer can be caused by direct UV radiation or by indirect stress via reactive oxygen species (ROS) [30]. The SNP rs142310826 identified in our study is about 1000kb upstream of the *NEIL3* gene, which encodes a DNA glycosylase that recognizes and removes lesions produced by oxidative stress, such as spiroiminodihydantoin (Sp), guanidinohydantoin (Gh), and 8-oxoguanine (8-oxoG), primarily in single-stranded DNA (ssDNA) [31]. This gene has been shown to be an important facilitator of cell proliferation in neural stem/progenitor cells and tumor cells, suggesting its possible role in replication-associated DNA repair [32-34]. Some studies have reported that polymorphisms of DNA glycosylases may

possess altered enzymatic activity, increasing the risk of inflammation-related cancers [35-37]. Our understanding of the mechanism of caffeine's inhibitory effect on BCC development is quite limited, though previous studies suggested a role of the ATR-Chk1 signaling pathway [16-19]. Similar to targets of NEIL3, the ATR binds to the chromosome at the site of ssDNA damage, which then leads to activation of checkpoints, DNA repair, and apoptosis to prevent damaged cells from progressing through the cell cycle [38]. Caffeine could either directly disrupt the ATR-Chk1 checkpoint pathway [19] or inhibit ATR-mediated DNA repair [39], and prematurely increase the number of cells that undergo apoptosis. In addition, caffeine has been proven to exert antioxidant effects that could neutralize oxidative stress in cells [40], which may decrease oxidative DNA damage and alter the expression of related DNA repair genes, such as *NEIL3*.

The genetic markers identified among women were mapped on genes *LRRTM4*, *ATF3*, and *DCLRE1C*. The leucine-rich repeat transmembrane neuronal 4 (LRRTM4) may play a role in the development and maintenance of the excitatory synapse in the vertebrate nervous system [41]. SNPs mapped on this gene are associated with phenotypes such as verbal declarative memory, sporadic amyotrophic lateral sclerosis, and immunoglobulin G (IgG) glycosylation in the GWAS catalog (http://www.ebi.ac.uk/gwas/home). Changes in IgG glycosylation have been linked to gastric cancer and ovarian cancer in previous studies [42, 43], indicating other possible functions of this gene. The activating transcription factor 3 (*ATF3*) gene has been demonstrated to play opposite roles (oncogene or tumor suppression) in cancer development depending on the cell type and context [44]. Though upregulation of ATF3 appears to enhance tumor formation in keratinocytes [45], Atf3 protein levels decreased when caffeine was administered in a mouse

model [46]. Moreover, ATF3 was discovered to be related to the ATR-Chk1 pathway as well [47]. The DNA cross-link repair 1C (*DCLRE1C*) gene encodes a nuclear protein that has single-strand-specific exonuclease activity and also functions in the regulation of the cell cycle in response to DNA damage [48].

SNPs that showed the greatest interaction with caffeine in BCC development in males were mapped on gene *POTEA*. *POTE* is a highly homologous gene family located on numerous chromosomes and expressed in a wide variety of human cancers (colon, lung, breast, ovary, and pancreas) [49]. In normal tissue, its expression is restricted to testis, ovary, and prostate, with the highest expression in testis [50]. Little is known about the biological function of this gene family, but there is evidence for its role in inducing programmed cell death [50].

We specifically extracted analysis results for caffeine consumption-related loci identified by previous GWAS analysis. None of them showed significant interaction with caffeine intake in relation to BCC risk. We additionally tested the individual and combined associations between caffeine SNPs and risk of BCC, but none reached statistical significance. These results suggest that the inverse association between caffeine intake and risk of BCC is not due to interaction or association with already-known caffeine-related loci.

Our study has several strengths. First, we used high-quality cohort data, among which information on both caffeine intake and genetic markers is available for studying G-E interactions. The relatively large sample size facilitated detection of potential interaction, even using a conventional logistic regression approach and a stringent genome-wide significance level. Second, we took gender difference into consideration in our analysis, because men and women

may have different caffeine consumption habits, and caffeine may interact with sex hormones when exerting its biological effects [51]. This study design helps us identify several loci that are specific to men or women, although further studies are needed to verify our findings. Third, we modeled caffeine consumption in a different way for the purpose of sensitivity analysis, and the results did not change materially.

We also acknowledge some limitations: First, we used caffeine consumption in 1986 rather than cumulative average intake in our analysis. To study G-E interaction, environmental exposure should be measured at appropriate time points, because many genes are expressed only during specific developmental periods, and some exposures may have greater impact on specific stages. However, we understand little about the biological mechanisms and induction period of caffeine's effects on the development of BCC. Given that the induction period of cancer is relatively long, and computing cumulative average caffeine intake is not easy among BCC controls, caffeine intake in 1986 is an acceptable option. Second, SNP rs142310826 is relatively rare and of only moderate imputation quality in our datasets. However, we directly genotyped this SNP among 335 participants in our GWAS datasets, the correlation between imputed dosage and directly genotyped allele count was 0.7 (p-value<0.0001). Third, because all the participants in the current study are health professionals of European ancestry, our results may not be generalizable to other ethnic or socioeconomic groups. Finally, we did not split our data into discovery and replication sets, because combined analysis across all studies is the most powerful analytical strategy [52].

In conclusion, in this genome-wide G-E interaction meta-analysis, the association of caffeine

intake with BCC risk differed according to genetic variation of SNP rs142310826. Genetic markers at chromosomes 2p12, 1q32.2, 10p13, and 8p11.21 modified the caffeine-BCC association differently in men and women. Further G-E interaction analyses are warranted to verify our findings, and additional biological studies are needed to better elucidate the roles of these genetic variants and their mapped genes.

Acknowledgements

We are indebted to the participants in the NHS, NHS2, and HPFS for their dedication to this research. We thank the following state cancer registries for their help: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington, and Wyoming. The authors assume full responsibility for analyses and interpretation of these data. This work is supported by NIH R01 CA49449, P01 CA87969, UM1 CA186107, and UM1 CA167552.

References

1. Wong, C., R. Strange, and J. Lear, *Basal cell carcinoma.* BMJ: British Medical Journal, 2003. **327**(7418): p. 794.
2. Diepgen, T. and V. Mahler, *The epidemiology of skin cancer.* British Journal of Dermatology, 2002. **146**(s61): p. 1-6.
3. Gallagher, R.P., et al., *Sunlight exposure, pigmentary factors, and risk of nonmelanocytic skin cancer: I. Basal cell carcinoma.* Archives of Dermatology, 1995. **131**(2): p. 157.
4. Lear, J., et al., *Risk factors for basal cell carcinoma in the UK: case-control study in 806 patients.* Journal of the Royal Society of Medicine, 1997. **90**(7): p. 371.
5. Han, J., G.A. Colditz, and D.J. Hunter, *Risk factors for skin cancers: a nested case–control study within the Nurses' Health Study.* International journal of epidemiology, 2006. **35**(6): p. 1514-1521.
6. Stacey, S.N., et al., *Common variants on 1p36 and 1q42 are associated with cutaneous basal cell carcinoma but not with melanoma or pigmentation traits.* Nature genetics, 2008. **40**(11): p. 1313-1318.
7. Stacey, S.N., et al., *New common variants affecting susceptibility to basal cell carcinoma.* Nature genetics, 2009. **41**(8): p. 909-914.
8. Nan, H., et al., *Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma.* Human molecular genetics, 2011. **20**(18): p. 3718-3724.
9. Lu, Y.-P., et al., *Topical applications of caffeine or (−)-epigallocatechin gallate (EGCG)*

inhibit carcinogenesis and selectively increase apoptosis in UVB-induced skin tumors in mice. Proceedings of the National Academy of Sciences, 2002. **99**(19): p. 12455-12460.

10. Kerzendorfer, C. and M. O'Driscoll, *UVB and caffeine: inhibiting the DNA damage response to protect against the adverse effects of UVB.* Journal of Investigative Dermatology, 2009. **129**(7): p. 1611-1613.

11. Lou, Y.-R., et al., *Effects of oral administration of tea, decaffeinated tea, and caffeine on the formation and growth of tumors in high-risk SKH-1 mice previously treated with ultraviolet B light.* Nutrition and cancer, 1999. **33**(2): p. 146-153.

12. Abel, E.L., et al., *Daily coffee consumption and prevalence of nonmelanoma skin cancer in Caucasian women.* european Journal of Cancer prevention, 2007. **16**(5): p. 446-452.

13. Rees, J.R., et al., *Tea consumption and basal cell and squamous cell skin cancer: results of a case-control study.* Journal of the American Academy of Dermatology, 2007. **56**(5): p. 781-785.

14. Stensvold, M.I. and B.K. Jacobsen, *Coffee and cancer: a prospective study of 43,000 Norwegian men and women.* Cancer Causes & Control, 1994. **5**(5): p. 401-408.

15. Song, F., A.A. Qureshi, and J. Han, *Increased caffeine intake is associated with reduced risk of basal cell carcinoma of the skin.* Cancer research, 2012. **72**(13): p. 3282-3289.

16. Han, W., M. Ming, and Y.-Y. He, *Caffeine promotes ultraviolet B-induced apoptosis in human keratinocytes without complete DNA repair.* Journal of Biological Chemistry, 2011. **286**(26): p. 22825-22832.

17. Heffernan, T.P., et al., *ATR–Chk1 pathway inhibition promotes apoptosis after UV*

treatment in primary human keratinocytes: Potential basis for the UV protective effects of caffeine. Journal of Investigative Dermatology, 2009. **129**(7): p. 1805-1815.

18. Kumagai, A., et al., *The Xenopus Chk1 protein kinase mediates a caffeine-sensitive pathway of checkpoint control in cell-free extracts.* The Journal of cell biology, 1998. **142**(6): p. 1559-1569.

19. Lu, Y.-P., et al., *Effect of caffeine on the ATR/Chk1 pathway in the epidermis of UVB-irradiated mice.* Cancer research, 2008. **68**(7): p. 2523-2529.

20. Rimm, E.B., et al., *Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals.* American Journal of Epidemiology, 1992. **135**(10): p. 1114-1126.

21. Feskanich, D., et al., *Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire.* Journal of the American Dietetic Association, 1993. **93**(7): p. 790-796.

22. Colditz, G.A., et al., *Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women.* American Journal of Epidemiology, 1986. **123**(5): p. 894-900.

23. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.

24. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-2191.

25. Aulchenko, Y.S., M.V. Struchalin, and C.M. van Duijn, *ProbABEL package for*

genome-wide association analysis of imputed data. BMC bioinformatics, 2010. **11**(1): p. 134.

26. Sulem, P., et al., *Sequence variants at CYP1A1–CYP1A2 and AHR associate with coffee consumption.* Human molecular genetics, 2011. **20**(10): p. 2071-2077.

27. Amin, N., et al., *Genome-wide association analysis of coffee drinking suggests association with CYP1A1/CYP1A2 and NRCAM.* Molecular psychiatry, 2011. **17**(11): p. 1116-1129.

28. Cornelis, M.C., et al., *Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption.* PLoS genetics, 2011. **7**(4): p. e1002033.

29. Cornelis, M., et al., *Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption.* Molecular psychiatry, 2014.

30. Ichihashi, M., et al., *UV-induced skin damage.* Toxicology, 2003. **189**(1): p. 21-39.

31. Krokeide, S.Z., et al., *Human NEIL3 is mainly a monofunctional DNA glycosylase removing spiroimindiohydantoin and guanidinohydantoin.* DNA repair, 2013. **12**(12): p. 1159-1164.

32. Morland, I., et al., *Human DNA glycosylases of the bacterial Fpg/MutM superfamily: an alternative pathway for the repair of 8‐oxoguanine and other oxidation products in DNA.* Nucleic acids research, 2002. **30**(22): p. 4926-4936.

33. Liu, M., et al., *Expression and purification of active mouse and human NEIL3 proteins.* Protein expression and purification, 2012. **84**(1): p. 130-139.

34. Liu, M., S. Doublié, and S.S. Wallace, *Neil3, the final frontier for the DNA glycosylases that recognize oxidative damage.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2013. **743**: p. 4-11.

35. Nohmi, T., S.-R. Kim, and M. Yamada, *Modulation of oxidative mutagenesis and carcinogenesis by polymorphic forms of human DNA repair enzymes.* Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2005. **591**(1): p. 60-73.

36. Chen, L., et al., *Association between polymorphism of human oxoguanine glycosylase 1 and risk of prostate cancer.* The Journal of urology, 2003. **170**(6): p. 2471-2474.

37. Tudek, B., *Base excision repair modulation as a risk factor for human cancers.* Molecular aspects of medicine, 2007. **28**(3): p. 258-275.

38. Batista, L.F., et al., *How DNA lesions are turned into powerful killing structures: insights from UV-induced apoptosis.* Mutation Research/Reviews in Mutation Research, 2009. **681**(2): p. 197-208.

39. Selby, C.P. and A. Sancar, *Molecular mechanisms of DNA repair inhibition by caffeine.* Proceedings of the National Academy of Sciences, 1990. **87**(9): p. 3522-3525.

40. Devasagayam, T., et al., *Caffeine as an antioxidant: inhibition of lipid peroxidation induced by reactive oxygen species.* Biochimica et Biophysica Acta (BBA)-Biomembranes, 1996. **1282**(1): p. 63-70.

41. Song, Y.S. and E. Kim, *Presynaptic proteoglycans: sweet organizers of synapse development.* Neuron, 2013. **79**(4): p. 609-611.

42. Saldova, R., et al., *Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG.* Glycobiology, 2007. **17**(12): p. 1344-1356.

43. Ruhaak, L.R., et al., *The serum immunoglobulin G glycosylation signature of gastric cancer.* EuPA open proteomics, 2015. **6**: p. 1-9.

44. Thompson, M.R., D. Xu, and B.R. Williams, *ATF3 transcription factor and its emerging roles in immunity and cancer.* Journal of molecular medicine, 2009. **87**(11): p. 1053-1060.

45. Wu, X., et al., *Opposing roles for calcineurin and ATF3 in squamous skin cancer.* Nature, 2010. **465**(7296): p. 368-372.

46. Jia, H., et al., *Coffee intake mitigated inflammation and obesity-induced insulin resistance in skeletal muscle of high-fat diet-induced obese mice.* Genes & nutrition, 2014. **9**(3): p. 1-10.

47. Demidova, A.R., et al., *Dual regulation of Cdc25A by Chk1 and p53-ATF3 in DNA replication checkpoint control.* Journal of Biological Chemistry, 2009. **284**(7): p. 4132-4139.

48. Felgentreff, K., et al., *Functional analysis of naturally occurring DCLRE1C mutations and correlation with the clinical phenotype of ARTEMIS deficiency.* Journal of Allergy and Clinical Immunology, 2015.

49. Bera, T.K., et al., *POTE paralogs are induced and differentially expressed in many cancers.* Cancer research, 2006. **66**(1): p. 52-56.

50. Liu, X.F., et al., *A primate-specific POTE-actin fusion protein plays a role in apoptosis.* Apoptosis, 2009. **14**(10): p. 1237-1244.

51. Ascherio, A., et al., *Caffeine, postmenopausal estrogen, and risk of Parkinson's disease.* Neurology, 2003. **60**(5): p. 790-795.

52. Skol, A.D., et al., *Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.* Nature genetics, 2006. **38**(2): p. 209-213.

Supplementary Materials

1. Detailed description of study population and GWAS sets

Nurses' Health Study (NHS): The NHS is a prospective cohort study established in 1976 with 121,700 female U.S registered nurses, who were then 30-55 years old. All of them completed and returned a mailed self-administered questionnaire about their medical histories and lifestyle at the baseline. In 1989 and 1990, a total of 32,826 women provided blood samples. Information regarding medical history, lifestyle, and disease diagnoses was updated every 2 years with a follow-up rate of 90%.

Health Professionals Follow-up Study (HPFS): The HPFS began in 1986 with 51,529 U.S. male health professionals who were 40-75 years old at initial recruitment. They all answered a detailed mailed questionnaire at the inception of the study. Disease- and health-related information was obtained and updated through biennial questionnaires. Between 1993 and 1994, 18,159 of these men provided a blood sample. The average follow-up rate for this cohort over 10 years is greater than 90%.

Supplementary Table 2.1 Basic information on the 18 GWAS sets from NHS and HPFS

Study	Sample size * (Genotyped)	Genotyping platform	Combined dataset
Postmenopausal invasive breast cancer case-control study nested within the NHS (NHS-BrCa)	1145 cases, 1142 controls	Illumina 550k	Illumina
Type 2 diabetes case-control study nested within the NHS (NHS-T2D)	1532 cases, 1754 controls	Affy 6.0	Affy

Supplementary Table 2.1 Basic information on the 18 GWAS sets from NHS and HPFS (Continued)

Study	Sample size * (Genotyped)	Genotyping platform	Combined dataset
Coronary heart disease case-control study nested within the NHS (NHS-CHD)	342 cases, 804 controls	Affy 6.0	Affy
Kidney stone case-control study nested within the NHS (NHS-KS)	328 cases, 166 controls	Illumina 610Q	Illumina
Pancreas cancer case-control study nested within the NHS (NHS- Pancreas)	82 cases, 84 controls	Illumina 550k	Illumina
Glaucoma case-control study nested within the NHS (NHS-Glaucoma)	313 cases, 497 controls	Illumina 660	Illumina
Endometrial cancer case-control study nested within the NHS (NHS-Endometrial)	396 cases, 348 controls	Omni Express	Omni
Colon cancer case-control study nested within the NHS (NHS-Colon)	394 cases, 774 controls	Omni Express	Omni
Mammographic density study nested within the NHS (NHS-Mammographic density)	153 cases, 641 controls	Omni Express	Omni
Gout case-control study nested within the NHS (NHS-Gout)	319 cases, 392 controls	Omni Express	Omni
Type 2 diabetes case-control study nested within the HPFS (HPFS-T2D)	1189 cases, 1298 controls	Affy 6.0	Affy
Coronary heart disease case-control study nested within the HPFS (HPFS-CHD)	435 cases, 878 controls	Affy 6.0	Affy
Kidney stone case-control study nested within the HPFS (HPFS-KS)	315 cases, 238 controls	Illumina 610Q	Illumina
Pancreas cancer case-control study nested within the HPFS (HPFS-Pancreas)	54 cases, 52 controls	Illumina 550k	Illumina
Advanced prostate cancer case-control study nested within the HPFS (HPFS-AdvPrCa)	218 cases, 205 controls	Illumina 610Q	Illumina
Glaucoma case-control study nested within the HPFS (HPFS-Glaucoma)	178 cases, 299 controls	Illumina 660	Illumina
Colon cancer case-control study nested within the HPFS (HPFS-Colon)	229 cases, 230 controls	Omni Express	Omni
Gout case-control study nested within the HPFS (HPFS-Gout)	717 cases, 699 controls	Omni Express	Omni

* These are number of participants who have been genotyped in each of the studies before imputation, quality control, and further exclusion. Cases refer to the cases of disease in the original nested case-control study.

2. Genotyping, quality control, and imputation

Genotyping

There were 18 GWAS datasets from the NHS and HPFS with cleaned genotype data available. We combined these datasets into three complied datasets based on their genotype platform type: Affymetrix (Affy), Illumina HumanHap series (Illumina), or Illumina Omni Express (Omni). The Affymetrix dataset was comprised of data on the Affy 6.0 platform (NHS-type 2 diabetes, NHS-coronary heart disease, HPFS-type 2 diabetes, HPFS-coronary heart disease). The Illumia HumanHap dataset was comprised of several platforms: Illumina 550K (NHS-breast cancer, NHS-Pancreas cancer, HPFS-pancreas cancer), Illumina 610Q (NHS-kidney stone, HPFS-kidney stone, HPFS-prostate cancer) and Illumina 660 (NHS-glaucoma, HPFS-glaucoma). The Illumina Omni Express dataset contained only studies genotyped on the Omni Express platform (NHS-endometrial cancer, NHS-colon cancer, NHS-mammographic density, NHS-gout, HPFS-colon, HPFS-gout). Detailed method about the pooled imputed data in this combined dataset is described in Lindström, et al. submitted to Bioinformatics (copy is provided for reviewers' review).

Quality control (QC)

We combined the individual datasets that were genotyped on the same platform, removing any SNPs that were not in all studies and with a missing call rate>5%, and flipping strands where appropriate to create a final compiled dataset. This resulted in 668,283 SNPs in the Affymetrix

dataset, 459,999 SNPs in the Illumina HumanHap dataset, and 565,810 SNPs in the Illumina Omni Express dataset. Analyses were restricted to subjects with self-reported European ancestry. Genetic principal components were calculated using sets of independent SNPs (12,000-33,000 SNPs depending on platform). Subjects who did not cluster with other self-identified Europeans based on the top five principal components were also excluded.

We then ran a pairwise identity by descent (IBD) analysis for each combined dataset to detect duplicate and related individuals based on resulting Z scores. If 0<=Z0<=0.1 and 0<=Z1<=0.1 and 0.9 <=Z2<=1.1 then a pair was flagged as being identical twins or duplicates. Pairs were considered full siblings if 0.17<=Z0<=0.33 and 0.4 <=Z1<=0.6 and 0.17<=Z2<=0.33. Half siblings or avunculars were defined as having 0.4<=Z1<=0.6 and 0<=Z2<=0.1. Some of the duplicates flagged in this step were expected, having been genotyped in multiple datasets and hence having the same cohort IDs. In this case, one of each pair was randomly chosen for removal from the dataset. Instances where pairs were flagged as unexpected duplicates with the different cohort IDs, but pairwise genotype concordance rate>0.999, resulted in removal of both individuals from the pair. Related individuals (full sibs, half sibs/avunculars) were not removed from the final datasets. In the Affymetrix dataset, 167 individuals were removed because they were duplicates or were flagged for removal from secondary genotype data cleaning, leaving a total of 8065 individuals. Of the 6894 individuals originally in the Illumina dataset, 107 were removed because they were duplicates or flagged for removal in the genotyping step, leaving 6787 IDs. In addition, 8 pairs of individuals were flagged as related. In the Omni express dataset, there were 5956 individuals at the start, with 39 IDs to remove leaving 5917 IDs and 5 pairs of

related IDs.

After removing duplicate IDs and flagging related pairs of IDs, we used EIGENSTRAT [1] to run PCA analysis on each compiled dataset, removing one member from each flagged pair of related individuals. For Affymetrix and Illumina HumanHap, we used approximately 12,000 SNPs that were filtered to ensure low pairwise LD. For the OmniExpress dataset we used approximately 33,000 SNPs that were similarly filtered. We plotted the top eigenvectors using R and examined the plots for outliers.

Finally as a quality control check, we ran logistic regression analyses using each individual study's controls as "cases" and the rest of the studies controls as "controls". We then ran regressions with each of the other study controls as "cases" versus all of the rest of the controls. We looked for p values of genome-wide significance ($p<10^{-8}$) and examined QQ plots to determine if any SNPs were flagged as significant where no SNPs should have been significant. In the Affymetrix dataset 100 SNPs were flagged and removed. In the Illumina HumanHap dataset, 8 SNPs had $p<10^{-8}$ in any of the QC regressions and were removed. No SNPs in the Illumina Omni Express dataset had p values$<10^{-8}$, hence no additional SNPs needed to be removed. After the datasets were combined and appropriate SNP and ID filters applied, the complied datasets were imputed.

Imputation

After the datasets were combined and appropriate quality control procedures applied, the

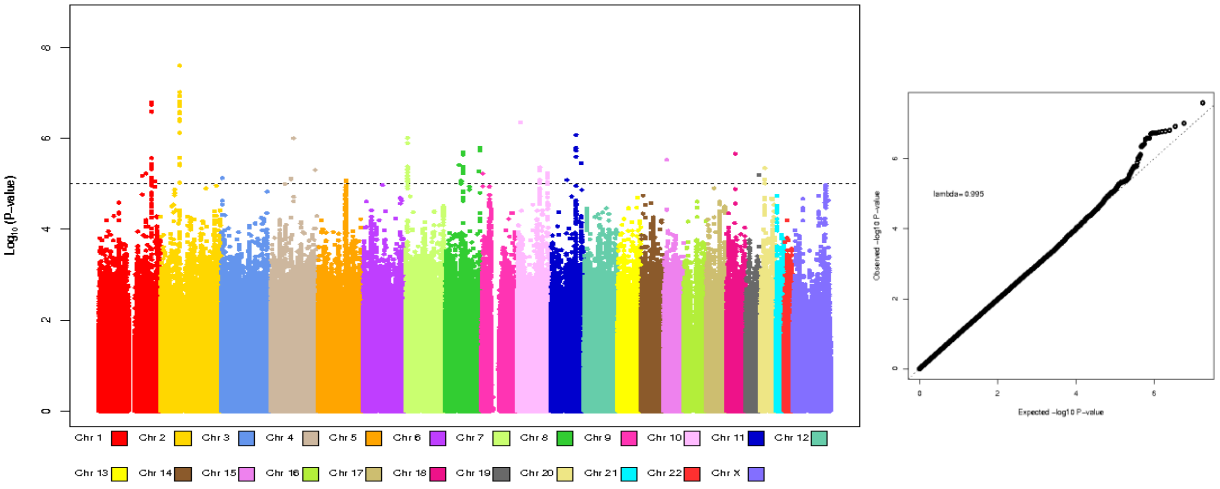
complied datasets were imputed using the 1000 Genomes Project ALL Phase I Integrated Release Version 3 Haplotypes excluding monomorphic and singleton sites (2010-11 data freeze, 2012-03-14 haplotypes) as reference panel. SNP genotypes were imputed in three steps. First, genotypes on each chromosome were split into chunks to facilitate windowed imputation in parallel using ChunkChromosome (http://genome.sph.umich.edu/wiki/ChunkChromosome, v. 2011-08-05). Then each chunk of chromosome was phased using MACH (v. 1.0.18.c) [2]. In the final step, Minimac (v. 2012-08-15) [3] was used to impute the phased genotypes to approximately 31 million markers in the 1000 Genomes Project. The number of genotyped SNPs passed quality control procedure and that of imputed SNPs with minor allele frequency (MAF) > 1% and imputation R^2>0.3 in each platform are presented in ***Supplementary Table 2.2***.

Supplementary Table 2.2 Summary of markers in combined datasets

Platform	# of markers in cleaned and merged datasets	Total # of 1000G imputed markers	# of 1000G imputed markers with MAF>1%	# of 1000G imputed markers with MAF>1% and imputation R^2> 0.3
Affymetrix (Affy)	668,283	31,326,389	9,783,513	9,783,513
Illumina (Illumina)	459,999	31,326,389	9,807,739	8,991,321
Omni Express (Omni)	565,810	31,326,389	9,771,868	9,148,255

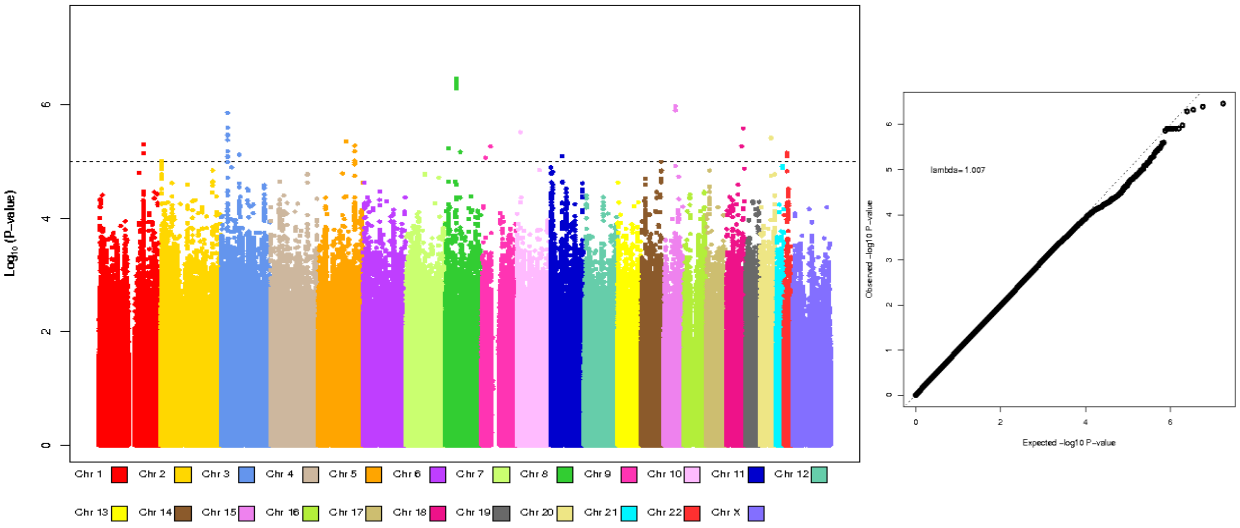
3. Supplementary Results

Supplementary Figure 2.1 Manhattan plot and Q-Q plot for the interaction results in the NHS &&



&& Results are for the meta-analysis of Illumina_NHS, Affy_NHS, and Omni_NHS. Quartiles of caffeine intake were modeled as an ordinal variable (main analysis).

Supplementary Figure 2.2 Manhattan plot and Q-Q plot for the interaction results in the HPFS



&& Results are for the meta-analysis of Illumina_HPFS, Affy_HPFS, and Omni_HPFS. Quartiles of caffeine intake were modeled as an ordinal variable (main analysis).

Supplementary Table 2.3 Genetic markers with p-value for interaction < 5×10^{-6} in the NHS

Marker Name(CHR:BP)	Allele 1	Allele 2	Freq1 %	Effect	StdErr	P-value	P-Het &
1:212702723	c	g	0.553	0.1761	0.0375	2.72E-06	0.8121
1:212708134	t	c	0.1272	0.2554	0.0558	4.80E-06	0.4712
1:212709301	t	c	0.1273	0.2553	0.0558	4.68E-06	0.4704
1:212709466	t	c	0.8727	-0.2554	0.0558	4.63E-06	0.4707
1:212710947	t	c	0.1273	0.2557	0.0557	4.34E-06	0.4725
1:212711790	t	c	0.1273	0.256	0.0556	4.22E-06	0.4717
1:212715522	a	g	0.8726	-0.2571	0.0556	3.77E-06	0.4644
1:212716491	t	c	0.8726	-0.2573	0.0556	3.68E-06	0.4627
1:212721010	t	c	0.8537	-0.2733	0.0531	2.62E-07	0.4975
1:212721216	t	c	0.8556	-0.2782	0.0533	1.82E-07	0.4491
1:212724028	a	g	0.1442	0.2801	0.0536	1.74E-07	0.4524
1:212726219	a	g	0.1511	0.2702	0.0524	2.55E-07	0.3471
1:212728988	a	g	0.1528	0.2739	0.0523	1.64E-07	0.2726
1:212729555	a	g	0.1527	0.2738	0.0523	1.67E-07	0.2704
10:15021717	a	t	0.0105	-1.3215	0.2619	4.51E-07	0.9034
10:92155938:A_AT	i	r	0.3394	-0.1787	0.0389	4.34E-06	0.8978
11:101777151	a	g	0.4049	0.1838	0.0373	8.38E-07	0.3157
11:101778710	a	g	0.5922	-0.1769	0.0369	1.67E-06	0.2897
11:101779050	a	g	0.5922	-0.1768	0.0369	1.69E-06	0.2891
11:101781044	a	g	0.4103	0.1781	0.0371	1.61E-06	0.2841
11:101781634	a	g	0.4078	0.1764	0.037	1.87E-06	0.2929
11:101785248	a	g	0.4092	0.1749	0.0372	2.53E-06	0.2647
11:101787112	a	g	0.5473	-0.1819	0.038	1.69E-06	0.3556
11:122841232:G_GCT	i	r	0.361	0.2103	0.0454	3.55E-06	0.5426
15:35117508	a	g	0.9013	0.3883	0.0831	2.97E-06	0.437
18:35818913	a	g	0.3123	0.1932	0.0408	2.19E-06	0.7619
2:76670131	t	c	0.1891	-0.2209	0.0479	3.93E-06	0.9366
2:76671644	a	g	0.1886	-0.2212	0.0477	3.62E-06	0.9343
2:76673246	t	c	0.189	-0.2239	0.0477	2.66E-06	0.9252
2:76676942	a	g	0.1828	-0.2369	0.0479	7.57E-07	0.934
2:76692117	a	g	0.1828	-0.2428	0.0479	4.04E-07	0.9414
2:76701736	a	t	0.8155	0.241	0.0476	4.16E-07	0.9956
2:76702062	t	g	0.8152	0.2419	0.0476	3.78E-07	0.9932
2:76704590	t	c	0.8146	0.25	0.0476	1.53E-07	0.9946

Supplementary Table 2.3 Genetic markers with p-value for interaction < 5×10^{-6} in the NHS (Continued)

Marker Name(CHR:BP)	Allele 1	Allele 2	Freq1 %	Effect	StdErr	P-value	P-Het &
2:76711828	t	c	0.1835	-0.2542	0.048	1.18E-07	0.9518
2:76714536	c	g	0.8143	0.2481	0.0475	1.80E-07	0.9935
2:76714567	a	g	0.8143	0.248	0.0475	1.81E-07	0.9935
2:76722019	a	g	0.1831	-0.2492	0.0479	1.93E-07	0.9046
2:76722305	t	c	0.814	0.2436	0.0474	2.79E-07	0.9714
2:76725088	a	g	0.8141	0.2443	0.0474	2.60E-07	0.9778
2:76736544	a	t	0.8186	0.2585	0.0485	9.62E-08	0.9789
2:76738900:TTAGA	d	r	0.186	-0.2695	0.0484	2.51E-08	0.9626
20:17953062	a	c	0.4782	0.17	0.0371	4.53E-06	0.2878
4:179402856	a	t	0.0189	-0.816	0.1787	4.96E-06	0.7157
4:92729586	t	c	0.4311	0.182	0.0372	9.94E-07	0.3107
7:7337795	t	c	0.145	-0.2679	0.0547	9.86E-07	0.979
7:7337811	a	g	0.1482	-0.2613	0.054	1.28E-06	0.9674
7:7338285	a	g	0.8386	0.2313	0.0505	4.65E-06	0.9408
7:7338384	a	g	0.8385	0.2313	0.0505	4.68E-06	0.9392
7:7338620	t	c	0.1614	-0.2311	0.0506	4.93E-06	0.9227
7:7339392	a	t	0.1617	-0.2327	0.0505	4.15E-06	0.9276
7:7339551	t	c	0.1596	-0.2335	0.0511	4.81E-06	0.9159
8:138317880	a	g	0.2928	-0.1899	0.0398	1.85E-06	0.3364
8:138320626	a	t	0.2927	-0.1919	0.04	1.59E-06	0.3455
8:54802780	t	c	0.0282	0.6935	0.1501	3.82E-06	0.6769
8:70921095	a	c	0.826	-0.2217	0.0483	4.41E-06	0.8813
8:70922437	a	t	0.1744	0.2214	0.0483	4.60E-06	0.8901
8:70922442	t	c	0.1744	0.2214	0.0483	4.57E-06	0.8899
8:70923078	a	g	0.8269	-0.2223	0.0487	5.00E-06	0.9117
8:70923158	t	c	0.1745	0.2217	0.0483	4.46E-06	0.8896
8:70923473	t	c	0.8255	-0.2218	0.0483	4.43E-06	0.8901
8:70924230	t	c	0.1661	0.2371	0.0499	2.03E-06	0.9183
8:70924250	a	g	0.8291	-0.2316	0.049	2.29E-06	0.8879
8:70924383	t	g	0.8258	-0.222	0.0483	4.30E-06	0.8865
8:70924408	t	c	0.8258	-0.2223	0.0483	4.19E-06	0.8867

% Frequency for allele 1;

& P for heterogeneity comparing estimates of Illumina_NHS, Affy_NHS, and Omni_NHS.

Supplementary Table 2.4 Genetic markers with p-value for interaction < 5×10^{-6} in the HPFS

Marker Name (CHR:BP)	Allele 1	Allele 2	Freq1 %	Effect	StdErr	P-value	P-Het &
1:181588271	t	c	0.5176	0.1861	0.0408	4.99E-06	0.2566
10:13948900	t	c	0.011	-1.4326	0.3068	3.02E-06	0.5707
15:69384548	t	c	0.9751	-0.6513	0.1343	1.25E-06	0.07403
15:69385013	t	c	0.0249	0.651	0.1342	1.23E-06	0.07386
15:69385115	a	g	0.0243	0.6516	0.1343	1.23E-06	0.07547
15:69385597	t	c	0.0242	0.6499	0.134	1.23E-06	0.07438
15:69385651	t	c	0.9758	-0.6499	0.134	1.23E-06	0.07433
15:69385656	a	c	0.9758	-0.6493	0.1339	1.23E-06	0.07398
15:69386084	t	g	0.9758	-0.648	0.1327	1.05E-06	0.07713
18:68730423	c	g	0.0171	0.9784	0.2081	2.59E-06	0.2683
20:42631393	a	g	0.0151	0.9413	0.2038	3.87E-06	0.3307
20:42631701	a	t	0.985	-0.9439	0.2042	3.79E-06	0.3
3:25535771	t	c	0.7393	-0.2303	0.0477	1.38E-06	0.2116
3:25538317	a	c	0.2565	0.2243	0.0477	2.53E-06	0.2146
3:25538410	t	c	0.7355	-0.2174	0.0472	4.15E-06	0.3173
3:25538769:T_TCA	i	r	0.2655	0.2188	0.0471	3.43E-06	0.3184
3:25538883	a	g	0.7358	-0.2201	0.0474	3.36E-06	0.2976
5:111931224	t	c	0.9738	0.9367	0.204	4.38E-06	0.1724
8:43273165	a	g	0.9635	0.8859	0.1759	4.72E-07	0.7391
8:43285229	c	g	0.0343	-0.9745	0.1911	3.43E-07	0.8255
8:43290783	c	g	0.9625	0.8849	0.1746	4.01E-07	0.7656
8:43359009	t	c	0.0362	-0.8988	0.179	5.13E-07	0.7534

% Frequency for allele 1;

& P for heterogeneity comparing estimates of Illumina_HPFS, Affy_HPFS, and Omni_HPFS.

Supplementary Table 2.5 Genetic markers with p-value for interaction < 5×10^{-6} in the Meta-analysis of all datasets

Marker Name (CHR:BP)	Allele 1	Allele 2	Freq1 %	Effect	StdErr	P-value	P-Het &
1:61302504	a	g	0.9538	0.6127	0.1247	8.97E-07	0.6522
13:52427466	a	c	0.4913	-0.1259	0.0274	4.25E-06	0.4445

Supplementary Table 2.5 Genetic markers with p-value for interaction < 5×10^{-6} in the Meta-analysis of all datasets (Continued)

Marker Name (CHR:BP)	Allele 1	Allele 2	Freq1 %	Effect	StdErr	P-value	P-Het &
13:52433564	t	c	0.5306	0.1248	0.0268	3.12E-06	0.3477
14:100987542	c	g	0.0432	-0.5406	0.1089	6.91E-07	0.8224
14:101000359	a	g	0.0181	-0.6936	0.1447	1.64E-06	0.9337
14:33202232	t	c	0.9077	0.2298	0.0485	2.16E-06	0.8696
14:33203029	c	g	0.0899	-0.2267	0.0491	3.92E-06	0.7261
14:33204319	a	g	0.9118	0.2314	0.05	3.62E-06	0.6745
14:33204393	a	t	0.9112	0.228	0.0498	4.60E-06	0.7242
15:82348040:AGAT_	d	r	0.7053	-0.1495	0.0323	3.78E-06	0.4163
17:74634684	t	c	0.128	-0.193	0.0413	2.98E-06	0.2779
17:74636253	t	c	0.128	-0.1926	0.0413	3.09E-06	0.2759
17:74636448	c	g	0.8721	0.1925	0.0413	3.13E-06	0.277
17:74637002	a	g	0.1296	-0.1902	0.0412	3.94E-06	0.2651
17:74637022	t	c	0.872	0.191	0.0413	3.67E-06	0.2798
17:74637143	t	c	0.8719	0.1903	0.0412	3.93E-06	0.2787
17:74637472	a	t	0.8719	0.1901	0.0412	4.01E-06	0.2794
17:74637649	t	c	0.8718	0.19	0.0412	3.99E-06	0.2943
17:74637738	c	g	0.8717	0.1904	0.0412	3.81E-06	0.297
17:74637881	a	g	0.8717	0.1904	0.0412	3.82E-06	0.2975
17:74638033	a	g	0.1278	-0.19	0.0414	4.35E-06	0.2891
17:74638036	t	c	0.8721	0.1899	0.0413	4.25E-06	0.2797
17:74638043	t	g	0.1279	-0.1899	0.0413	4.25E-06	0.2797
17:74638485	t	c	0.8719	0.1896	0.0412	4.21E-06	0.2831
17:74638873	a	g	0.1281	-0.1895	0.0412	4.30E-06	0.2886
2:223558675	a	g	0.0936	-0.2507	0.0535	2.85E-06	0.1861
4:179402856	a	t	0.0185	-0.7571	0.1344	1.78E-08	0.6168
4:92704420	a	g	0.6245	-0.1372	0.0287	1.78E-06	0.2636
4:92720480	a	g	0.6205	-0.1319	0.0278	2.00E-06	0.2829
6:149913630	a	t	0.9407	-0.486	0.104	2.97E-06	0.228
7:138696170	t	g	0.014	0.7159	0.1555	4.13E-06	0.6094
7:138710477	a	t	0.0152	0.742	0.1541	1.48E-06	0.4866
7:67820903	a	g	0.0728	-0.2483	0.0528	2.54E-06	0.3462
8:110870998	t	c	0.025	-0.4426	0.0963	4.31E-06	0.9386
8:110876861	t	g	0.0269	-0.4829	0.1017	2.05E-06	0.7887

% Frequency for allele 1;

& P for heterogeneity between results of NHS and HPFS (gender difference).

Supplementary Table 2.6 Sensitivity analysis [++] results for the independent markers identified in the main analysis

Rs number – Identified Group [a]	CHR:BP	NHS [b]		HPFS [c]		All [d]	
		Beta	P-value	Beta	P-value	Beta	P-value
NA [&] – NHS	2:76738900:TTAGA	-0.0013	1.59E-07	-0.0002	0.45	NA	NA
rs77868414 – HPFS	8:43285229	0.0003	0.70	-0.0052	4.95E-07	NA	NA
rs142310826 - ALL	4:179402856	-0.0045	6.26E-06	-0.0033	2.02E-03	-0.0040	6.10E-08

a Identified group = NHS if the genetic marker's *p* for interaction is less than 5×10^{-7} among NHS participants; = HPFS if the marker was identified among HPFS participants; =ALL if identified among all participants;

b Results of meta-analysis of Illumina_NHS, Affy_NHS, and Omni_NHS;

c Results of meta-analysis of Illumina_HPFS, Affy_HPFS, and Omni_HPFS;

d Results of meta-analysis of Illumina_NHS, Affy_NHS, Omni_NHS, Illumina_HPFS, Affy_HPFS, and Omni_HPFS; NA means the estimates in men and women are significantly different.

++ In sensitivity analysis, we used the median values of each quartile to represent the corresponding caffeine intake levels (1st quartile = median intake of 1st quartile, 2nd quartile=median intake of 2nd quartile, etc.)

& This marker does not have a rs number

References

1. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.

2. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genetic epidemiology, 2010. **34**(8): p. 816-834.

3. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nature genetics, 2012. **44**(8): p. 955-959.

CHAPTER 3

Height, height-related SNPs, and risk of non-melanoma skin cancer

Xin Li [1], Liming Liang [1], Immaculata De Vivo [1,2], Edward Giovannucci [1,2,3], Jean Y. Tang [4], and Jiali Han [5]

[1] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

[2] Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA;

[3] Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

[4] Department of Dermatology, Stanford University School of Medicine, Redwood City, CA, USA;

[5] Department of Epidemiology, Fairbanks School of Public Health, Indiana University, and Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, IN, USA

Abstract

Adult height, determined by genetics and by early-life exposures, has been associated with an increased risk of several site-specific cancers, including skin cancer. However, less attention has been given to non-melanoma skin cancer (NMSC). Using the data from the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS), we prospectively examined the risk of squamous cell carcinoma (SCC) and basal cell carcinoma (BCC) in relation to adult height. After controlling for potential confounding factors, the hazard ratios (HRs) were 1.09 (95% CI: 1.03, 1.16) and 1.10 (95% CI: 1.07, 1.12) for the associations between every 10cm increase in height and risk of SCC and BCC respectively. No significant interaction between height and other risk factors was observed. In the genetic datasets of the NHS and HPFS, none of the 687 height-related SNPs was significantly associated with risk of SCC or BCC after Bonferroni correction. The associations between genetic scores combining independent height-related loci and NMSC risk were not significant either. Our data from two large cohorts provide further evidence that height is associated with increased risk of non-melanoma skin cancer. More studies on height-related genetic loci and early-life exposures may help clarify the underlying mechanisms.

Introduction

Non-melanoma skin cancer (NMSC), including basal and squamous cell carcinomas (BCC and SCC, respectively), is the most common malignancy among white people [1]. It is estimated that over 2 million cases of NMSC occur each year in the US, with the incidence continues to increase [2]. BCC rarely metastasizes to other organs or causes death; however this malignancy results in considerable morbidity and places a huge burden on healthcare system worldwide [3]. In contrast, SCC is more likely to invade other tissues and could lead to death [3]. Both environmental and constitutional factors contribute to the development of NMSC. Ultraviolet radiation is a well-established carcinogen for both BCC and SCC [4, 5]. Constitutional risk factors that represent certain components of genetic susceptibility include hair color, family history, tanning ability, and so forth. [6-8].

Taller people are more likely to develop cancer [9]. Though a number of case-control [10-13] and cohort studies [14-18] have examined the association between adult height and risk of melanoma skin cancer, the association between height and risk of NMSC has been sparsely investigated. One prospective study reported a significant higher risk of NMSC among taller men and women [16]. However, BCC and SCC were not analyzed separately. Besides, the authors failed to consider important confounders such as race, constitutional factors and sun exposure history, as well as potential effect modifications by them. Therefore, a comprehensive assessment of the relationships between height and risk of different types of NMSC is still lacking.

The underlying mechanism for this positive association remains unclear. One possible explanation is that height-related genetic factors are also tied to skin cancer; however studies

exploring this possibility are rare. Adult height is determined by genetic factors to a great extent [19]. The largest genome-wide association study (n = 253,288) on height has been conducted by the Genetic Investigation of Anthropometric Traits (GIANT) consortium, in which they identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height [20]. Testing the associations between these height-related single-nucleotide polymorphisms (SNPs) and NMSC risk may help better understand the relationship between these two phenotypes and provide more insight into skin tumorigenesis

In the present study, we used data from the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) to investigate the association between height and risk of incident SCC and BCC simultaneously. We also evaluated the extent to which the observed associations were affected by confounding factors, and tested potential interactions between height and other factors on NMSC risk. In order to better understand the association at the genetic level, we also examined the individual and combined associations of height-related variants identified by the GIANT consortium with risk of NMSC in the genetic datasets of the NHS and HPFS.

Methods

Study Population

Nurses' Health Study (NHS): The NHS is a prospective cohort study established in 1976 with 121,700 female U.S registered nurses, who were then 30-55 years old. All of them completed and returned a mailed self-administered questionnaire about their medical histories

and baseline lifestyle. In 1989 and 1990, a total of 32,826 women provided blood samples. Information regarding medical history, lifestyle, and disease diagnoses was updated every two years with a follow-up rate of 90%.

Health Professionals Follow-up Study (HPFS): The HPFS began in 1986 with 51,529 U.S. male health professionals who were 40-75 years old at initial recruitment. They all answered a detailed mailed questionnaire at the inception of the study. Disease- and health-related information was obtained and updated through biennial questionnaires. Between 1993 and 1994, 18,159 of these men provided a blood sample. The average follow-up rate for this cohort over 10 years is greater than 90%.

Genetic datasets: Eighteen case-control studies nested within the NHS and HPFS with cleaned genotype data were included in our study. Samples from the 18 studies were genotyped using a variety of platforms, which we then combined into three compiled datasets based on their genotype platform types: Affymetrix (Affy), Illumina HumanHap series (Illumina), or Illumina Omni Express (Omni). Quality control on SNP completion rate, sample completion rate, ancestry consistency, deviation from Hardy-Weinberg equilibrium (HWE), Mendelian consistency, minor allele frequency, and duplication were conducted within each of the three combined datasets. We then imputed the compiled datasets using the 1000 Genomes Project ALL Phase I Integrated Release Version 3 Haplotypes excluding monomorphic and singleton sites (2010-11 data freeze, 2012-03-14 haplotypes) as the reference panel. Basic information on the 18 studies and detailed descriptions of quality control and imputation are provided in ***Supplementary Materials***.

Measurement of height and ascertainment of skin cancers

Height was reported by participants at recruitment (1976 for NHS, 1986 for HPFS). New diagnoses of non-melanoma skin cancer were reported by participants biennially. With their permission, participants' medical records were obtained and reviewed by physicians to confirm the diagnoses of SCC. Though medical records were not obtained for BCC, the validity of BCC self-reports was more than 90% in our study [21, 22].

Measurement of covariates

Information on skin cancer risk factors was obtained from questionnaires in both the NHS and the HPFS in the 1980s. The risk factors included: (1) natural hair color at age 20; (2) family history of melanoma in first-degree relatives; (3) skin reaction after 2 hours of sun exposure as a child/adolescent; (4) number of severe sunburns over lifetime; (5) mole count measuring 3mm or larger on the left arm; and (6) states lived in at birth, age 15, and age 30.

Data on weight, smoking status, and menopausal status was first collected at baseline (1976 for NHS and 1986 for HPFS) and then updated biennially in subsequent questionnaires for all cohort members. Body mass index (BMI) was computed as weight in kilograms divided by the square of height in meters for each follow-up cycle. Physical activity was first asked with detail in 1986 in both cohorts and updated every two years thereafter. The reproducibility and validity of self-reported physical activity in both cohorts has been evaluated in detail in previous studies [23, 24]. Energy expenditure in metabolic equivalent tasks (METs) [25] measured in hours per week was calculated by multiplying the number of hours per week of leisure-time physical

activity by the metabolic equivalent (MET) value of the activity and summing the products of all types of activities. Food frequency questionnaires were initially collected in 1980 for the NHS and 1986 for the HPFS, and alcohol intake and diet were generally updated every four years. Previous studies have shown that the food-frequency questionnaire validly assesses dietary and alcohol intake during the past year [26, 27]. Self-reported race which was measured in 1982 in NHS and 1986 in HPFS was also considered. Non-whites were collapsed into one group because of insufficient sample sizes in individual race categories.

Height-related SNPs and calculation of genetic score

Of 697 height-related SNPs identified by the GIANT consortium, 687 were available in our genetic dataset. For a locus in which multiple SNPs in linkage disequilibrium (LD, defined as $r^2 > 0.1$) were identified, we selected the SNP with the most significant association with height as reported by the GIANT paper, leaving 593 SNPs for genetic score calculation. The scores were calculated only for individuals who had no missing value in any of the chosen SNPs. We assumed an additive genetic model for each SNP, which performs well even when the true genetic model is unknown or wrongly specified [28]. For each individual, we summed the dosage of alleles that are related to increase in height of those independent loci to obtain the simple count genetic score. We also constructed a weighted score by multiplying the dosage of effect alleles by the corresponding regression coefficients in the original GWAS paper and then summing the products. Both the original simple count score and the weighted score were rescaled to a mean of 1186 alleles (2 alleles * 593 SNPs) before testing their associations with

NMSC to make the results comparable.

Statistical analysis

Height and skin cancer: Participants who did not report their date of birth or height were excluded, as were those who had invalid information on height at recruitment (whose reported height was < 120 or >200 cm). Participants who had baseline cancers were excluded, and those who reported any type of cancer or died during follow-up were also excluded from subsequent follow-up. We used cox proportional hazards models stratified by follow-up cycles and age to calculate the hazard ratios (HRs) and 95% confidence intervals (CIs) of each type of skin cancer. Person-time was calculated for each participant from the date of baseline questionnaire return to the date of the first report of NMSC, death, or the end of follow-up (June 2010), whichever came first. When quantifying the relationship between NMSC and height, we modeled height as a continuous measure expressed in 10cm (increasing) increments. In the multivariate analysis, we simultaneously controlled for age, smoking status, alcohol intake, BMI, physical activity, and menopausal status/postmenopausal hormone use (only in NHS). Then, we fitted a more complex model by additionally including hair color, family history of melanoma, sunburn reaction as a child/adolescent, number of severe sunburns, mole count, and states lived in at birth, age 15, and age 30. Lastly, race was controlled for in the model to assess potential confounding effects. We tested the heterogeneity of the results obtained among men and women and conducted a meta-analysis if there was no significant gender difference. Multiplicative interactions between height and other potential risk factors of NMSC were tested by using the likelihood ratio test

comparing a "main effect only" model vs. a model with the product term. All covariates in the multivariable-adjusted models were considered and sequentially tested for interaction each at a time. All statistical analyses were performed using SAS software (version 9.3 for UNIX; SAS Institute, Cary, North Carolina). We considered 2-sided P values less than 0.05 to be statistically significant.

Height-related SNPs and skin cancer: Data on participants who appeared in more than one of the three combined datasets were included only once in analyses. Baseline common cancer cases were excluded, as were NMSC cases who had other common cancers before diagnosis of skin cancer. Eligible controls were free of skin cancers or other common cancers. We assessed the associations between individual height-related SNPs and SCC as well as BCC using logistic regression models adjusted for gender, age, and the top three eigenvectors (EVs). The same models were fitted for the associations between genetic scores and risk of NMSC. All the analyses were first conducted within each of the platform-specific datasets, and then combined by meta-analysis if results were not significantly different. ProbABEL package and R-3.0.2 were used to perform these tests. We considered 2-sided P values less than 0.05 to be statistically significant. Bonferroni correction using the number of independent tests was applied to account for multiple comparisons.

Sensitivity analysis and validation of self-reported ancestry

Ancestry within the white population is a potential confounder that may bias the estimation of height-skin cancer association. Height varies across Europe, with Northern Europeans

generally taller than Southern Europeans [29-31]. Intra-European ethnic origin has also been found to be related to both melanoma and non-melanoma skin cancers [32, 33]. We adjusted self-reported race (Southern European/Mediterranean; Scandinavian; Other Caucasian; and None-white ancestry) in the multivariable models for height-NMSC association; however, such information may be inaccurate.

Therefore, we used participants' genetic data to estimate their accurate ancestry. Genetic ancestry was represented by ancestry coordinates that were calculated by the Locating Ancestry from Sequence Reads (LASER) method. This method has been demonstrated to accurately infer worldwide continental ancestry and even the fine-scale ancestry within Europe. Detailed descriptions of LASER have been published previously [34, 35]. We tested the correlation between self-reported European ancestry and the first as well as the second ancestry coordinates to validate the information collected by the questionnaire. We also conducted a sensitivity analysis in which we compared the cox models without ancestry, with self-reported ancestry, and with genetic ancestry coordinates as covariates. These analyses were restricted to participants in the genetic dataset, all of whom are of European ancestry.

Results

Height and skin cancer risk

We included 117,887 and 50,767 participants from the NHS and the HPFS, respectively. We documented 1,646 SCCs over 3,198,317 person-years and 18,681 BCCs during 3,187,992 person-years in the NHS. In the HPFS, 1,244 SCC events during 862,935 person-years and 9,625

BCCs over 854,657 person-years of follow-up were identified.

The baseline age-standardized characteristics of participants by quartiles of height are listed in ***Table 3.1***. Taller participants tended to be younger, drank more alcohol, excised more, and were more likely to be current smokers. Higher prevalence of Southern European ethnicity, family history of melanoma, red/blond hair, presence of arm moles, and painful burn/blister skin reaction after prolonged sun exposure as a child/adolescent were found in higher quartiles of height. Study participants with short stature had a higher BMI than taller participants. These trends were consistent in men and women. In the NHS, the percentage of current hormone replacement therapy (HRT) users is higher among taller women.

In the age-adjusted models (Model 1 in ***Table 3.2***) and multivariate models without race (Model 2 & 3), height was significantly positively associated with risk of SCC and BCC in both men and women. Further including self-reported race (Model 4) did not alter the results materially in the NHS. Risk of SCC only showed a borderline association with height in the HPFS. In the full model (Model 4), HRs for the associations between per 10 cm increase in height and SCC were 1.09 (95% CI: 1.01, 1.19) in women and 1.09 (95% CI: 1.00, 1.19) in men. For BCC, the HRs were 1.11 (95% CI: 1.08, 1.13) and 1.08 (95% CI: 1.05, 1.12), respectively, among females and males. Though the magnitude of association measures appeared to be slightly higher in the NHS, heterogeneity between genders did not reach statistical significance (P for het = 0.91, and 0.23 for SCC and BCC, respectively) (***Table 3.2***). We found no significant interaction between height and other covariates in the full multivariable-adjusted model.

Height-related SNPs and skin cancer risk

Sample sizes of the genetic datasets before exclusion and number of NMSC cases and controls after exclusion are shown in ***Table 3.3***. Among the 687 height-related SNPs available in our genetic dataset, 37 and 38 showed nominally significant associations (P value < 0.05) with risk of SCC and BCC respectively (***Supplementary Tables 3.3 & 3.4***). However, none of them was significantly associated with risk of skin cancers after Bonferroni correction. Mean values and ranges of the genetic scores combining all 593 independent (R^2 for LD < 0.1) height-related SNPs were similar among the Illumina, Affy, and Omni datasets (***Table 3.4a & 3.4b***). The genetic scores were significantly associated with height in our genetic datasets. However, we observed no significant association between the scores and risk of NMSC (***Table 3.4a & 3.4b***). We constructed two types of genetic scores using the formula shown in ***Table 3.4***. The results for simple count score and weighted score were similar to each other.

Sensitivity analysis and validation of self-reported ancestry

The Pearson correlations between self-reported European ancestry and the first ancestry coordinate were 0.23, 0.28, and 0.31 in Affy, Illumina, and Omni datasets, respectively (all P-values < 0.0001). The Pearson correlations between self-reported European ancestry and the second ancestry coordinate were -0.14, -0.16, and -0.16 in Affy, Illumina, and Omni datasets, respectively (all P-values < 0.0001). Results of the multivariable-adjusted models with self-reported ancestry and the models with genetic ancestry were not different materially (***Supplementary Table 3.5***).

Table 3.1 Baseline characteristics by quartiles of height in the NHS (1976-2010) and HPFS (1986-2010)

	Quartiles of height in cm							
	NHS (women)				HPFS (men)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Age (year) [a], mean (SD)	42.8(7.3)	42.5(7.3)	42.2(7.2)	41.7(7.1)	56.3(10.2)	54.7(9.9)	53.8(9.6)	52.3(9.3)
Self-reported Race								
- Southern European/Mediterranean, %	21.7	18.2	16.5	14.2	29.0	25.5	23.7	21.0
- Scandinavian, %	4.4	6.2	7.2	9.0	7.6	9.6	12.7	13.6
- Other Caucasian, %	53.2	57.6	59.1	60.1	55.0	61.1	60.5	62.7
- Non-white, %	20.7	17.9	17.2	16.8	8.4	3.9	3.1	2.6
Family history of melanoma, %	6.8	7.1	7.4	7.6	4.0	3.8	4.0	4.0
Red/blonde hair, %	13.4	15.1	16.1	17.6	11.1	14.3	14.7	15.0
Presence of arm moles, %	33.7	36.2	37.2	39.2	30.6	32.8	32.7	33.4
Painful burn/blisters reaction as a child/adolescent, %	14.0	14.5	14.4	15.7	23.4	24.6	24.6	25.8
Number of blistering sun burn >=5, %	6.4	7.5	7.4	8.3	31.6	34.8	35.3	37.4
Current smoking, %	32.1	32.8	33.5	35.2	8.8	9.2	9.3	9.4
Alcohol intake (g/d), mean (SD)	5.8(10.6)	6.5(11.0)	6.8(11.3)	7.2(11.7)	10.5(15.2)	11.8(16.0)	12.1(16.4)	12.8(17.2)
Body mass index (kg/m2), mean (SD)	24.0(4.3)	23.9(4.2)	23.6(4.1)	23.5(4.0)	25.7(3.7)	25.4(3.1)	25.4(3.1)	25.5(3.1)
Physical activity (metabolic-equivalents hours/wk), mean (SD)	14.0(20.6)	13.8(20.6)	14.3(21.3)	14.2(21.5)	20.5(29.9)	20.6(28.4)	20.9(30.9)	21.3(30.4)
Menopausal status/PMH status								
- Premenopausal, %	80.5	80.5	80.5	80.3	NA			
- HRT never use, %	9.2	9.0	9.3	9.0				
- HRT current use, %	6.5	6.9	6.8	7.2				
- HRT past use, %	3.8	3.7	3.4	3.6				

NOTE: Values are means (SD) or percentages and are standardized to the age distribution of the study populations. Values of multi-level categorical variables may not sum to 100% due to rounding. HRT stands for hormone replacement therapy

a Value is not age-adjusted.

Table 3.2 HRs and 95% CIs for the associations of height (per 10cm increase) with SCC and BCC risk

	NHS		HPFS		Meta-analysis		
	HR (95% CI)	**P-value**	**HR (95% CI)**	**P-value**	**HR (95% CI)**	**P-value**	**P for Het #**
SCC							
- Model 1	1.19 (1.10, 1.28)	<.0001	1.18 (1.09, 1.28)	0.0001	1.18 (1.12, 1.25)	<.0001	0.9263
- Model 2	1.16 (1.07, 1.26)	0.0002	1.16 (1.07, 1.26)	0.0006	1.16 (1.10, 1.23)	<.0001	0.9821
- Model 3	1.11 (1.02, 1.20)	0.0125	1.11 (1.01, 1.21)	0.0227	1.11 (1.04, 1.17)	0.0007	0.9864
- Model 4	1.09 (1.01, 1.19)	0.0272	1.09 (1.00, 1.19)	0.0610	1.09 (1.03, 1.16)	0.0038	0.9081
BCC							
- Model 1	1.17 (1.15, 1.20)	<.0001	1.14 (1.11, 1.18)	<.0001	1.16 (1.13, 1.19)	<.0001	0.1416
- Model 2	1.16 (1.13, 1.18)	<.0001	1.13 (1.09, 1.16)	<.0001	1.14 (1.11, 1.17)	<.0001	0.2146
- Model 3	1.12 (1.09, 1.14)	<.0001	1.10 (1.06, 1.13)	<.0001	1.11 (1.09, 1.13)	<.0001	0.3862
- Model 4	1.11 (1.08, 1.13)	<.0001	1.08 (1.05, 1.12)	<.0001	1.10 (1.07, 1.12)	<.0001	0.2349

a Model 1, age-adjusted;

b Model 2 adjusted for age, smoking status (never, past, current 1-14,15-24, or 25+ cigarettes/day), alcohol intake (no, <5.0, 5.0-9.9, 10.0-19.9, or 20.0+ g/day), body mass index (<25.0, 25.0-29.9, 30.0-34.9, or 35.0+ kg/m^2), physical activity (<3.0, 3.0-8.9, 9.0-17.9, 18.0-26.9 or 27.0+ metabolic equivalent hours/wk), menopausal status/postmenopausal hormones use (premenopausal, HRT never, HRT past, or HRT current; only in the NHS);

c Model 3 adjusted for covariates in Model 2, plus natural hair color (red, blonde, light brown, dark brown, or black), childhood/adolescent sunburn reaction (none or some redness, burn, painful burn or blisters), family history of melanoma (yes or no), number of severe sunburns over life time (never, 1-2 times, 3-5 times, or 6+ times), mole count (none, 1-2, 3-9, 10+), and states lived at birth, age 15, and age 30 (UV index <=5, =6, or >=7);

d Model 4 adjusted for covariates in Model 3 and race (Southern European/Mediterranean, Scandinavian, other Caucasian, or other race group);

Tests for heterogeneity between the NHS and HPFS (gender differences).

Table 3.3 Sample size of each platform-specific dataset before exclusion; Number of NMSC cases and controls [#] in each of the combined datasets after exclusion

Dataset	Sample size before exclusion	BCC		SCC	
		cases	controls	cases	controls
Affy	8065	1781	4304	247	4500
Illumina	5222	1055	1929	134	2029
Omni	5253	1062	2297	146	2433
Total	18540	3898	8530	527	8962

1976 and 1986 were considered baseline years for the NHS and the HPFS respectively. Skin cancer cases who had diagnosis of other common cancers before diagnosis of skin cancers were excluded; controls who had other cancers were excluded; participants with identical genetic information but different cohort ID were removed; participants sampled in more than one study were included only once. Participants who withdrew consent were excluded.

Table 3.4a Association between simple count genetic score of height-related SNPs and risk of NMSC [&]

	Illumina			Affy			Omni			Meta-analysis [%]		
Original Mean(range)	591.4 (538.6, 651.3)			591.0 (535.1, 650.6)			590.9 (537.4, 644.7)			--		
Formula	$2.0054(\sum_{i=1}^{593} SNP_i)$			$2.0067(\sum_{i=1}^{593} SNP_i)$			$2.0074(\sum_{i=1}^{593} SNP_i)$					
Rescaled Mean(range)	1186 (1080, 1306)			1186 (1074, 1306)			1186 (1079, 1294)					
	Beta	**SE**	**P-value**	**Beta**	**SE**	**P-value**	**Beta**	**SE**	**P-value**	**OR (95% CI) or Beta**	**P-value**	**P Het**
SCC	-0.005	0.003	0.130	0.0001	0.002	0.952	0.0008	0.003	0.783	1.00 (1.00, 1.00)	0.5317	0.296
BCC	-0.001	0.001	0.571	-0.0002	0.001	0.867	-0.002	0.001	0.177	1.00 (1.00, 1.00)	0.0647	0.443
Height	0.024	0.003	3.99E-16	0.016	0.002	4.74E-12	0.019	0.003	5.83E-10	0.02	<0.0001	0.084

& Logistic regression models were used to assess the relationship between genetic score of height SNPs and risk of skin cancers, adjusting for age, gender, and the top three eigenvectors.

% Analyses were first conducted within each of the platform-specific genetic datasets. We used fixed-effect meta-analysis to obtain a combined estimation. P Het is p-value for heterogeneity comparing three combined datasets: Illumina, Affy, and Omni.

Table 3.4b Association between weighted genetic score of height-related SNPs and risk of NMSC [&]

	Illumina			Affy			Omni			Meta-analysis [%]		
Original Mean(range)	17.17 (15.28, 18.65)			17.16 (15.48, 18.82)			17.16 (15.66, 18.73)			--		
Formula	$69.0739(\sum_{i=1}^{593}\beta_i SNP_i)$			$69.1142(\sum_{i=1}^{593}\beta_i SNP_i)$			$69.1142(\sum_{i=1}^{593}\beta_i SNP_i)$					
Rescaled Mean(range)	1186 (1055, 1288)			1186 (1070, 1300)			1186 (1082, 1295)					
	Beta	**SE**	**P-value**	**Beta**	**SE**	**P-value**	**Beta**	**SE**	**P-value**	**OR (95% CI) or Beta**	**P-value**	**P Het**
SCC	-0.005	0.003	0.115	0.0006	0.002	0.800	0.003	0.003	0.377	1.00 (1.00, 1.00)	0.8793	0.145
BCC	-0.001	0.001	0.399	-0.0003	0.001	0.785	-0.002	0.001	0.122	1.00 (1.00, 1.00)	0.0567	0.482
Height (cm)	0.027	0.003	<2E-16	0.018	0.002	6.03E-14	0.022	0.003	3.55E-12	0.02	<0.0001	0.042

& Logistic regression models were used to assess the relationship between genetic score of height SNPs and risk of skin cancers, adjusting for age, gender, and the top three eigenvectors.

% Analyses were first conducted within each of the platform-specific genetic datasets. We used fixed-effect meta-analysis to obtain a combined estimation. P Het is p-value for heterogeneity comparing three combined datasets: Illumina, Affy, and Omni.

Discussion

In this analysis of two large and well-characterized cohorts, height was positively associated with risk of both SCC and BCC. To assess confounding due to potential factors, we fitted three multivariable models and gradually added covariates. The magnitude of associations changed the most when skin cancer constitutional factors and sunburns were adjusted for. Self-reported race did not alter the estimates materially when other covariates were already in the models. The multivariable-adjusted HRs for BCC risk among women were greater than the corresponding ones among men, though tests of gender difference did not yield any significant findings. Thus, we combined the estimates of two cohorts by fixed-effect meta-analysis. The combined HRs were 1.09 (95% CI: 1.03, 1.16) and 1.10 (95% CI: 1.07, 1.12) for the associations of each 10cm increase in height with risk of SCC and BCC, respectively. There were much fewer events for SCC than for BCC over the follow-up period, thus the confidence intervals for the former were wider.

The most important non-genetic factors affecting height are nutritional status, living conditions, and serious disease during childhood/adolescence [36]. Height could thus be thought of as a marker for these early-life exposures rather than a risk factor itself. Both animal studies and epidemiological studies have shown that reduced caloric intake during development reduces future risk of malignancy [37-39]. Attention has also focused on the potential mechanistic relevance of growth factors and hormones. Higher levels of circulating insulin-like growth factor promote linear growth during childhood and has been shown to accelerate cell proliferation [40] and to inhibit apoptosis [41]. Another possible explanation is that height may be associated with

greater skin surface area, which may put more skin cells at risk of malignant transformation and progression to skin cancer [42].

Genetic factors contribute strongly to adult height. It has been estimated that 80% of the variation in height in Western populations is determined by genetics [43]. Some have proposed that the association between height and cancers may result from shared genetic components. Certain genes linked with height are also related to cancer regulatory pathways such as p53 and HH/PTCH [44]. Besides, height-related SNPs reported by the GIANT consortium have also been associated with risk of testicular cancer and prostate cancer [20]. Yet, it remains unclear whether these height SNPs are tied to skin cancer risk, individually or jointly. In our study, none of the 687 height-related SNPs was significantly associated with SCC or BCC risk after correcting for multiple comparisons. The genetic scores combining all independent SNPs showed no significant association with risk of SCC or BCC. It is possible that we are lacking power in our genetic datasets to observe the true associations.

The strengths of the current study include prospective design with long-term follow-up and high follow-up rate, availability of detailed information on a wide variety of covariates, involvement of both women and men, and targeting on SCC and BCC separately. A major advantage is that we examined the associations between height and skin cancers more thoroughly and accurately than has previously been reported. Potential confounding factors, such as pigmentation and sunburn history, which are critical for skin cancers and have not been considered before, were included in our cox models. We also took interactions between height and other covariates into consideration. In sensitivity analysis, ancestry within the white

population was assessed directly using genetic data. Adjustment for genetic ancestry did not change the results materially. This may result from lack of power in the genetic subsets and/or the control of skin cancer constitutional factors which have already partly explained variation in ancestry. Moreover, our novel analysis of the associations between height-related genetic variants and risk of skin cancers may eventually yield a better understanding of the underlying mechanisms. To our best knowledge, no such analysis has been conducted for skin cancers.

We also acknowledge several potential limitations of the present study. First, height was self-reported rather than measured in our cohorts, which could result in misclassification. However, any misclassification would be non-differential with respect to disease occurrence, because information on height was collected prior to the development of skin cancers. Because non-differential misclassification would bias the estimation downwards, that could not account for the observed positive association. Second, BCC cases were self-reported without further pathological confirmation. However, the high validity of self-reported BCC in these medically sophisticated populations has been confirmed in previous studies [22]. In addition, using the self-reported BCC cases, our group identified the previously well-documented genetic variant in the MC1R gene as the top risk locus in our GWAS for BCC [45]. These data support the validity of self-report of BCC in our study. Third, we did not have information on all relevant confounding variables. For example, data on socioeconomic status, which might affect both height and cancer incidence, were not available. However, our study used cohorts of health care providers, which has the advantage of minimizing confounding by educational attainment and adult socioeconomic status. In addition, adjustment for socioeconomic factors did not affect risk

estimates for association between height and cancer in previous large studies [15, 17, 46]. We also lacked information on childhood nutritional status, for which height may be a marker. Finally, our cohorts consist primarily of white health professionals and thus results may not be generalizable. However, such homogeneity in a study population would minimize confounding by socioeconomic status and differential access to healthcare and assure a high quality of returned data.

In conclusion, our data from two large cohorts provide further evidence that height is associated with increased risk of SCC and BCC. These associations were not explained by confounding by known risk factors, nor were modified by those risk factors. No significant association was observed between height-related genetic variants and risk of NMSC, no matter individually or jointly. More functional and epidemiological studies on height-related SNPs are needed to confirm our findings. Additional research involving a range of pre-adult exposures, such as diet, psychosocial stress, chronic illness, and social circumstances, which are rarely directly measured in existing datasets, may help clarify possible mechanisms underlying the positive associations.

References

1. Narayanan, D.L., R.N. Saladi, and J.L. Fox, *Review: Ultraviolet radiation and skin cancer.* International journal of dermatology, 2010. **49**(9): p. 978-986.
2. Rogers, H.W., et al., *Incidence estimate of nonmelanoma skin cancer in the United States, 2006.* Archives of dermatology, 2010. **146**(3): p. 283-287.
3. Tung, R.C. and A.T. Vidimos, *Non-melanoma skin cancer.* Retrieved March, 2002. **29**: p. 2007.
4. Gandini, S., et al., *Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure.* European Journal of Cancer, 2005. **41**(1): p. 45-60.
5. Armstrong, B.K., A. Kricker, and D.R. English, *Sun exposure and skin cancer.* Australasian Journal of Dermatology, 1997. **38**(S1): p. S1-S6.
6. Gandini, S., et al., *Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi.* European Journal of Cancer, 2005. **41**(1): p. 28-44.
7. Gandini, S., et al., *Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factors.* European Journal of Cancer, 2005. **41**(14): p. 2040-2059.
8. Han, J., G.A. Colditz, and D.J. Hunter, *Risk factors for skin cancers: a nested case–control study within the Nurses' Health Study.* International journal of epidemiology, 2006. **35**(6): p. 1514-1521.
9. Renehan, A.G., *Height and cancer: consistent links, but mechanisms unclear.* The lancet oncology, 2011. **12**(8): p. 716-717.

10. Cutler, C., et al., *Cutaneous malignant melanoma in women is uncommonly associated with a family history of melanoma in first-degree relatives: a case-control study.* Melanoma research, 1996. **6**(6): p. 435-440.

11. Gallus, S., et al., *Anthropometric measures and risk of cutaneous malignant melanoma: a case–control study from Italy.* Melanoma research, 2006. **16**(1): p. 83-87.

12. Olsen, C.M., et al., *Anthropometric factors and risk of melanoma in women: a pooled analysis.* International Journal of Cancer, 2008. **122**(5): p. 1100-1108.

13. Shors, A.R., et al., *Melanoma risk in relation to height, weight, and exercise (United States).* Cancer Causes & Control, 2001. **12**(7): p. 599-606.

14. Kabat, G.C., et al., *Adult stature and risk of cancer at different anatomic sites in a cohort of postmenopausal women.* Cancer Epidemiology Biomarkers & Prevention, 2013. **22**(8): p. 1353-1363.

15. Green, J., et al., *Height and cancer incidence in the Million Women Study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk.* The lancet oncology, 2011. **12**(8): p. 785-794.

16. Wirén, S., et al., *Pooled cohort study on height and risk of cancer and cancer death.* Cancer Causes & Control, 2014. **25**(2): p. 151-159.

17. Kabat, G.C., et al., *Adult height in relation to risk of cancer in a cohort of Canadian women.* International Journal of Cancer, 2013. **132**(5): p. 1125-1132.

18. Thune, I., et al., *Cutaneous malignant melanoma: Association with height, weight and body‐surface area. A prospective study in Norway.* International journal of cancer, 1993.

55(4): p. 555-561.

19. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nature genetics, 2010. **42**(7): p. 565-569.

20. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nature genetics, 2014. **46**(11): p. 1173-1186.

21. Hunter, D.J., et al., *Risk factors for basal cell carcinoma in a prospective cohort of women.* Annals of epidemiology, 1990. **1**(1): p. 13-23.

22. Colditz, G.A., et al., *Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women.* American Journal of Epidemiology, 1986. **123**(5): p. 894-900.

23. Wolf, A.M., et al., *Reproducibility and validity of a self-administered physical activity questionnaire.* International Journal of Epidemiology, 1994. **23**(5): p. 991-999.

24. Chasan-Taber, S., et al., *Reproducibility and validity of a self-administered physical activity questionnaire for male health professionals.* Epidemiology, 1996. **7**(1): p. 81-86.

25. Ainsworth, B.E., et al., *Compendium of physical activities: classification of energy costs of human physical activities.* Medicine & Science in Sports & Exercise, 1993(25): p. 71-80.

26. WILLETT, W.C., et al., *The use of a self-administered questionnaire to assess diet four years in the past.* American journal of epidemiology, 1988. **127**(1): p. 188-199.

27. SALVINI, S., et al., *Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption.* International journal of epidemiology, 1989.

18(4): p. 858-867.

28. Balding, D.J., *A tutorial on statistical methods for population association studies.* Nature Reviews Genetics, 2006. **7**(10): p. 781-791.

29. Grasgruber, P., et al., *The role of nutrition and genetics as key determinants of the positive height trend.* Economics & Human Biology, 2014. **15**: p. 81-100.

30. Turchin, M.C., et al., *Evidence of widespread selection on standing variation in Europe at height-associated SNPs.* Nature genetics, 2012. **44**(9): p. 1015-1019.

31. Cavelaars, A., et al., *Persistent variations in average height between countries and between socio-economic groups: an overview of 10 European countries.* Annals of human biology, 2000. **27**(4): p. 407-421.

32. D'Arcy, C., J. Holman, and B.K. Armstrong, *Pigmentary traits, ethnic origin, benign nevi, and family history as risk factors for cutaneous malignant melanoma.* Journal of the National Cancer Institute, 1984. **72**(2): p. 257-266.

33. English, D.R., et al., *Demographic characteristics, pigmentary and cutaneous risk factors for squamous cell carcinoma of the skin: A case‐control study.* International Journal of Cancer, 1998. **76**(5): p. 628-634.

34. Wang, C., et al., *Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation.* The American Journal of Human Genetics, 2015.

35. Wang, C., et al., *Ancestry estimation and control of population stratification for sequence-based association studies.* Nature genetics, 2014. **46**(4): p. 409-415.

36. Silventoinen, K., *Determinants of variation in adult body height.* Journal of biosocial science, 2003. **35**(02): p. 263-285.

37. Ross, M.H. and G. Bras, *Lasting influence of early caloric restriction on prevalence of neoplasms in the rat.* Journal of the National Cancer Institute, 1971. **47**(5): p. 1095-1114.

38. Ross, M.H. and G. Bras, *Tumor incidence patterns and nutrition in the rat.* The Journal of nutrition, 1965. **87**(3): p. 245-260.

39. Frankel, S., et al., *Childhood energy intake and adult mortality from cancer: the Boyd Orr Cohort Study.* Bmj, 1998. **316**(7130): p. 499-504.

40. Ish-Shalom, D., et al., *Mitogenic properties of insulin and insulin analogues mediated by the insulin receptor.* Diabetologia, 1997. **40**(2): p. S25-S31.

41. Milazzo, G., et al., *Insulin receptor expression and function in human breast cancer cell lines.* Cancer Research, 1992. **52**(14): p. 3924-3930.

42. Albanes, D. and M. Winick, *Are cell number and cell proliferation risk factors for cancer? 1.* Journal of the National Cancer Institute, 1988. **80**(10): p. 772-775.

43. Lai, C.-Q., *How much of human height is genetic and how much is due to nutrition.* Scientific Am, 2006.

44. Tripaldi, R., L. Stuppia, and S. Alberti, *Human height genes and cancer.* Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2013. **1836**(1): p. 27-41.

45. Nan, H., et al., *Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma.* Human molecular genetics, 2011. **20**(18): p. 3718-3724.

46. Sung, J., et al., *Height and site-specific cancer risk: a cohort study of a Korean adult population.* American journal of epidemiology, 2009: p. kwp088.

Supplementary Materials

1. Basic information on the 18 nested case-control studies

Supplementary Table 3.1 Basic information on the 18 GWAS sets from NHS and HPFS

Study	Sample size * (Genotyped)	Genotyping platform	Combined dataset
Postmenopausal invasive breast cancer case-control study nested within the NHS (NHS-BrCa)	1145 cases, 1142 controls	Illumina 550k	Illumina
Type 2 diabetes case-control study nested within the NHS (NHS-T2D)	1532 cases, 1754 controls	Affy 6.0	Affy
Coronary heart disease case-control study nested within the NHS (NHS-CHD)	342 cases, 804 controls	Affy 6.0	Affy
Kidney stone case-control study nested within the NHS (NHS-KS)	328 cases, 166 controls	Illumina 610Q	Illumina
Pancreas cancer case-control study nested within the NHS (NHS- Pancreas)	82 cases, 84 controls	Illumina 550k	Illumina
Glaucoma case-control study nested within the NHS (NHS-Glaucoma)	313 cases, 497 controls	Illumina 660	Illumina
Endometrial cancer case-control study nested within the NHS (NHS-Endometrial)	396 cases, 348 controls	Omni Express	Omni
Colon cancer case-control study nested within the NHS (NHS-Colon)	394 cases, 774 controls	Omni Express	Omni
Mammographic density study nested within the NHS (NHS-Mammographic density)	153 cases, 641 controls	Omni Express	Omni
Gout case-control study nested within the NHS (NHS-Gout)	319 cases, 392 controls	Omni Express	Omni
Type 2 diabetes case-control study nested within the HPFS (HPFS-T2D)	1189 cases, 1298 controls	Affy 6.0	Affy
Coronary heart disease case-control study nested within the HPFS (HPFS-CHD)	435 cases, 878 controls	Affy 6.0	Affy
Kidney stone case-control study nested within the HPFS (HPFS-KS)	315 cases, 238 controls	Illumina 610Q	Illumina
Pancreas cancer case-control study nested within the HPFS (HPFS-Pancreas)	54 cases, 52 controls	Illumina 550k	Illumina
Advanced prostate cancer case-control study nested within the HPFS (HPFS-AdvPrCa)	218 cases, 205 controls	Illumina 610Q	Illumina

Supplementary Table 3.1 Basic information on the 18 GWAS sets from NHS and HPFS (Continued)

Study	Sample size * (Genotyped)	Genotyping platform	Combined dataset
Glaucoma case-control study nested within the HPFS (HPFS-Glaucoma)	178 cases, 299 controls	Illumina 660	Illumina
Colon cancer case-control study nested within the HPFS (HPFS-Colon)	229 cases, 230 controls	Omni Express	Omni
Gout case-control study nested within the HPFS (HPFS-Gout)	717 cases, 699 controls	Omni Express	Omni

* These are number of participants who have been genotyped in each of the studies before imputation, quality control, and further exclusion. Cases refer to the cases of disease in the original nested case-control study.

2. Genotyping, quality control, and imputation

Genotyping

There were 18 GWAS datasets from the NHS and HPFS with cleaned genotype data available. We combined these datasets into three complied datasets based on their genotype platform type: Affymetrix (Affy), Illumina HumanHap series (Illumina), or Illumina Omni Express (Omni). The Affymetrix dataset was comprised of data on the Affy 6.0 platform (NHS-type 2 diabetes, NHS-coronary heart disease, HPFS-type 2 diabetes, HPFS-coronary heart disease). The Illumia HumanHap dataset was comprised of several platforms: Illumina 550K (NHS-breast cancer, NHS-Pancreas cancer, HPFS-pancreas cancer), Illumina 610Q (NHS-kidney stone, HPFS-kidney stone, HPFS-prostate cancer) and Illumina 660 (NHS-glaucoma, HPFS-glaucoma). The Illumina Omni Express dataset contained only studies genotyped on the

Omni Express platform (NHS-endometrial cancer, NHS-colon cancer, NHS-mammographic density, NHS-gout, HPFS-colon, HPFS-gout). Detailed method about the pooled imputed data in this combined dataset is described in Lindström, et al. submitted to Bioinformatics (copy is provided for reviewers' review).

Quality control (QC)

We combined the individual datasets that were genotyped on the same platform, removing any SNPs that were not in all studies and with a missing call rate>5%, and flipping strands where appropriate to create a final compiled dataset. This resulted in 668,283 SNPs in the Affymetrix dataset, 459,999 SNPs in the Illumina HumanHap dataset, and 565,810 SNPs in the Illumina Omni Express dataset. Analyses were restricted to subjects with self-reported European ancestry. Genetic principal components were calculated using sets of independent SNPs (12,000-33,000 SNPs depending on platform). Subjects who did not cluster with other self-identified Europeans based on the top five principal components were also excluded.

We then ran a pairwise identity by descent (IBD) analysis for each combined dataset to detect duplicate and related individuals based on resulting Z scores. If 0<=Z0<=0.1 and 0<=Z1<=0.1 and 0.9 <=Z2<=1.1 then a pair was flagged as being identical twins or duplicates. Pairs were considered full siblings if 0.17<=Z0<=0.33 and 0.4 <=Z1<=0.6 and 0.17<=Z2<=0.33. Half siblings or avunculars were defined as having 0.4<=Z1<=0.6 and 0<=Z2<=0.1. Some of the duplicates flagged in this step were expected, having been genotyped in multiple datasets and hence having the same cohort IDs. In this case, one of each pair was randomly chosen for

removal from the dataset. Instances where pairs were flagged as unexpected duplicates with the different cohort IDs, but pairwise genotype concordance rate>0.999, resulted in removal of both individuals from the pair. Related individuals (full sibs, half sibs/avunculars) were not removed from the final datasets. In the Affymetrix dataset, 167 individuals were removed because they were duplicates or were flagged for removal from secondary genotype data cleaning, leaving a total of 8065 individuals. Of the 6894 individuals originally in the Illumina dataset, 107 were removed because they were duplicates or flagged for removal in the genotyping step, leaving 6787 IDs. In addition, 8 pairs of individuals were flagged as related. In the Omni express dataset, there were 5956 individuals at the start, with 39 IDs to remove leaving 5917 IDs and 5 pairs of related IDs.

After removing duplicate IDs and flagging related pairs of IDs, we used EIGENSTRAT [1] to run PCA analysis on each compiled dataset, removing one member from each flagged pair of related individuals. For Affymetrix and Illumina HumanHap, we used approximately 12,000 SNPs that were filtered to ensure low pairwise LD. For the OmniExpress dataset we used approximately 33,000 SNPs that were similarly filtered. We plotted the top eigenvectors using R and examined the plots for outliers.

Finally as a quality control check, we ran logistic regression analyses using each individual study's controls as "cases" and the rest of the studies controls as "controls". We then ran regressions with each of the other study controls as "cases" versus all of the rest of the controls. We looked for p values of genome-wide significance ($p<10^{-8}$) and examined QQ plots to determine if any SNPs were flagged as significant where no SNPs should have been significant.

In the Affymetrix dataset 100 SNPs were flagged and removed. In the Illumina HumanHap dataset, 8 SNPs had $p<10^{-8}$ in any of the QC regressions and were removed. No SNPs in the Illumina Omni Express dataset had p values$<10^{-8}$, hence no additional SNPs needed to be removed. After the datasets were combined and appropriate SNP and ID filters applied, the complied datasets were imputed.

Imputation

After the datasets were combined and appropriate quality control procedures applied, the complied datasets were imputed using the 1000 Genomes Project ALL Phase I Integrated Release Version 3 Haplotypes excluding monomorphic and singleton sites (2010-11 data freeze, 2012-03-14 haplotypes) as reference panel. SNP genotypes were imputed in three steps. First, genotypes on each chromosome were split into chunks to facilitate windowed imputation in parallel using ChunkChromosome (http://genome.sph.umich.edu/wiki/ChunkChromosome, v. 2011-08-05). Then each chunk of chromosome was phased using MACH (v. 1.0.18.c) [2]. In the final step, Minimac (v. 2012-08-15) [3] was used to impute the phased genotypes to approximately 31 million markers in the 1000 Genomes Project. The number of genotyped SNPs passed quality control procedure and that of imputed SNPs with minor allele frequency (MAF) > 1% and imputation $R^2>0.3$ in each platform are presented in ***Supplementary Table 3.2***.

Supplementary Table 3.2 Summary of markers in combined datasets

Platform	# of markers in cleaned and merged datasets	Total # of 1000G imputed markers	# of 1000G imputed markers with MAF>1%	# of 1000G imputed markers with MAF>1% and imputation R^2> 0.3
Affymetrix (Affy)	668,283	31,326,389	9,783,513	9,783,513
Illumina (Illumina)	459,999	31,326,389	9,807,739	8,991,321
Omni Express (Omni)	565,810	31,326,389	9,771,868	9,148,255

3. Supplementary results

Supplementary Table 3.3 Height-related SNPs significantly associated with SCC risk at P-value <0.05

Marker Name (CHR: BP)	Allele 1	Allele 2	Freq 1	Effect	SE	P-value	P Het
13:50469913	t	c	0.20	0.27	0.08	9.40E-04	0.31
18:74983055	a	g	0.96	-0.46	0.14	1.01E-03	0.15
11:69163161	t	c	0.15	0.28	0.09	1.12E-03	0.25
5:172994624	a	g	0.35	-0.22	0.07	1.32E-03	0.76
16:990815	t	c	0.38	0.20	0.06	2.00E-03	0.89
9:119422807	a	t	0.92	-0.31	0.11	3.42E-03	0.42
7:73304636	t	c	0.14	0.25	0.09	4.59E-03	0.36
11:128577624	c	g	0.25	0.19	0.07	7.30E-03	0.20
11:45936035	a	g	0.92	0.37	0.14	8.31E-03	0.55
13:33143406	a	g	0.36	0.16	0.07	1.71E-02	0.42
8:135612595	a	g	0.30	0.16	0.07	1.78E-02	0.78
11:2171601	t	c	0.79	-0.19	0.08	2.01E-02	0.49
4:122720999	a	g	0.67	0.16	0.07	2.07E-02	0.13
9:18629792	a	g	0.31	0.16	0.07	2.49E-02	0.96
6:6889818	a	g	0.13	0.21	0.09	2.53E-02	0.23
15:94028149	t	c	0.64	0.15	0.07	2.55E-02	0.38

Supplementary Table 3.3 Height-related SNPs significantly associated with SCC risk at P-value <0.05 (Continued)

Marker Name (CHR: BP)	Allele 1	Allele 2	Freq 1	Effect	SE	P-value	P Het
13:30172751	t	g	0.21	0.17	0.08	2.65E-02	0.05
1:54954245	t	c	0.89	0.27	0.12	2.89E-02	0.48
5:171189571	a	g	0.67	0.15	0.07	3.05E-02	0.62
9:109518208	a	g	0.83	0.21	0.10	3.18E-02	0.94
19:19591066	a	g	0.83	0.20	0.09	3.19E-02	0.76
1:184007119	t	c	0.30	0.15	0.07	3.24E-02	0.17
12:124801226	t	c	0.27	-0.16	0.08	3.34E-02	0.64
17:54778817	a	t	0.67	0.15	0.07	3.39E-02	0.78
11:12678415	t	c	0.55	0.14	0.07	3.55E-02	0.97
1:67510474	a	g	0.28	-0.15	0.07	3.65E-02	0.73
4:184215675	a	g	0.23	-0.17	0.08	3.72E-02	0.24
2:44907331	a	g	0.14	-0.21	0.10	3.79E-02	0.79
10:127673877	t	c	0.33	0.14	0.07	3.79E-02	0.44
12:28952342	c	g	0.33	0.14	0.07	4.13E-02	0.42
6:45244415	t	c	0.58	-0.13	0.06	4.14E-02	0.18
2:233442091	a	g	0.28	-0.15	0.07	4.50E-02	0.23
9:117011595	t	c	0.37	-0.13	0.07	4.55E-02	0.35
8:145037573	a	t	0.40	0.13	0.06	4.58E-02	0.60
6:85448103	t	c	0.54	0.13	0.06	4.61E-02	0.55
6:105392745	t	c	0.69	0.14	0.07	4.95E-02	0.45
13:21570246	t	g	0.18	0.33	0.08	1.88E-05	0.97

NOTE: Freq 1 is the frequency of allele 1; P Het is p-value for heterogeneity comparing three combined datasets: Illumina, Affy, and Omni.

Supplementary Table 3.4 Height-related SNPs significantly associated with BCC risk at P-value <0.05

Marker Name (CHR: BP)	Allele 1	Allele 2	Freq 1	Effect	SE	P-value	P Het
7:150508720	t	c	0.29	0.11	0.03	2.58E-04	0.94
15:51269629	a	g	0.20	-0.12	0.04	7.10E-04	0.69
4:7055253	t	c	0.22	-0.11	0.03	1.03E-03	0.27
4:123835656	a	c	0.74	-0.10	0.03	1.70E-03	0.84

Supplementary Table 3.4 Height-related SNPs significantly associated with BCC risk at P-value <0.05 (Continued)

Marker Name (CHR: BP)	Allele 1	Allele 2	Freq 1	Effect	SE	P-value	P Het
14:55203126	c	g	0.76	-0.10	0.03	2.59E-03	0.96
15:89113138	c	g	0.33	-0.09	0.03	3.07E-03	0.39
3:68622366	a	g	0.55	-0.08	0.03	5.71E-03	0.96
5:176675423	a	g	0.98	0.26	0.10	1.01E-02	0.13
1:21583311	t	c	0.41	-0.07	0.03	1.20E-02	0.64
6:126216403	c	g	0.55	0.07	0.03	1.40E-02	0.23
3:190815978	a	g	0.89	0.11	0.05	1.76E-02	0.44
2:218284278	c	g	0.31	0.07	0.03	2.00E-02	0.23
3:129050943	a	g	0.21	-0.08	0.03	2.06E-02	0.35
14:65568215	a	g	0.37	-0.07	0.03	2.27E-02	0.62
2:242191410	c	g	0.75	0.07	0.03	2.29E-02	0.22
7:8086639	a	g	0.57	-0.07	0.03	2.56E-02	0.69
7:96039648	c	g	0.29	0.07	0.03	2.58E-02	0.82
4:8608634	t	c	0.43	-0.06	0.03	2.83E-02	0.05
1:172241251	t	c	0.68	0.07	0.03	2.85E-02	0.68
2:42462930	a	g	0.91	-0.11	0.05	2.86E-02	0.86
11:75276178	a	c	0.14	0.09	0.04	2.88E-02	0.49
7:23475919	a	g	0.14	0.09	0.04	3.28E-02	0.00
14:103878774	c	g	0.63	-0.06	0.03	3.32E-02	0.01
14:59688820	a	g	0.51	-0.06	0.03	3.32E-02	0.24
2:88924622	t	c	0.68	-0.06	0.03	3.33E-02	0.97
2:33315750	c	g	0.23	-0.07	0.04	3.68E-02	0.70
1:9292282	a	g	0.15	-0.08	0.04	4.07E-02	0.95
2:232779223	t	c	0.74	0.07	0.03	4.17E-02	0.94
5:88327782	t	g	0.52	0.06	0.03	4.29E-02	0.86
16:30030195	a	c	0.60	0.06	0.03	4.48E-02	0.60
6:144079629	t	g	0.34	-0.06	0.03	4.51E-02	0.91
10:105577409	a	g	0.10	0.09	0.05	4.55E-02	0.88
9:95387983	t	c	0.58	-0.06	0.03	4.55E-02	0.21
20:35544673	a	g	0.85	0.08	0.04	4.63E-02	0.26
7:46201355	a	g	0.78	0.07	0.03	4.66E-02	0.93
2:136187345	t	c	0.11	-0.09	0.05	4.76E-02	0.27
12:69827658	t	g	0.35	-0.06	0.03	4.85E-02	0.94
4:145565826	t	c	0.44	-0.06	0.03	4.93E-02	0.71

NOTE: Freq 1 is the frequency of allele 1; P Het is p-value for heterogeneity comparing three combined datasets: Illumina, Affy, and Omni.

Supplementary Table 3.5 HRs and 95% CIs for the associations of height (per 10cm increase) with SCC and BCC risk in sensitivity analysis

	NHS		HPFS	
	HR (95% CI)	**P-value**	**HR (95% CI)**	**P-value**
SCC				
- Model 1	1.24 (1.01, 1.51)	0.0366	1.01 (0.83, 1.23)	0.9292
- Model 2	1.21 (0.99, 1.48)	0.0624	0.99 (0.81, 1.21)	0.9108
- Model 3	1.16 (0.95, 1.42)	0.1415	0.93 (0.76, 1.14)	0.4952
- Model 4	1.14 (0.93, 1.40)	0.1957	0.92 (0.75, 1.13)	0.4452
- Model 5	1.14 (0.93, 1.39)	0.2221	0.92 (0.75, 1.13)	0.4221
BCC				
- Model 1	1.15 (1.08, 1.23)	<0.0001	1.14 (1.05, 1.23)	0.0017
- Model 2	1.13 (1.06, 1.21)	0.0003	1.12 (1.04, 1.22)	0.0043
- Model 3	1.09 (1.02, 1.17)	0.0098	1.10 (1.02, 1.20)	0.0178
- Model 4	1.09 (1.02, 1.17)	0.0117	1.10 (1.01, 1.19)	0.0220
- Model 5	1.08 (1.01, 1.16)	0.0185	1.12 (1.03, 1.21)	0.0078

a Model 1, age-adjusted

b Model 2 adjusted for adjusted for age, smoking status (never, past, current 1-14,15-24, or 25+ cigarettes/day), alcohol intake (no, <5.0, 5.0-9.9, 10.0-19.9, or 20.0+ g/day), body mass index (<25.0, 25.0-29.9, 30.0-34.9, or +35.0 kg/m^2), physical activity (<3.0, 3.0-8.9, 9.0-17.9, 18.0-26.9 or +27.0 metabolic equivalent hours/wk), menopausal status/postmenopausal hormones use (premenopausal, HRT never, HRT past, or HRT current; only in the NHS);

c Model 3 adjusted for covariates in Model 2, plus natural hair color (red, blonde, light brown, dark brown, or black), childhood/adolescent sunburn reaction (none or some redness, burn, painful burn or blisters), family history of melanoma (yes or no), number of severe sunburns over life time (never, 1-2 times, 3-5 times, or 6+ times), mole count (none, 1-2, 3-9, 10+), and states lived at birth, age 15, and age 30 (UV index <=5, =6, or >=7);

d Model 4 adjusted for covariates in Model 3 and self-reported white ancestry (Southern European/Mediterranean, Scandinavian, or other Caucasian);

e Model 5 adjusted for covariates in Model 3 and genetic ancestry (top 2 coordinates);

References

1. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.
2. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genetic epidemiology, 2010. **34**(8): p. 816-834.
3. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nature genetics, 2012. **44**(8): p. 955-959.