Title
Integrating Genetics Into Population-Based Studies

Author
Tai, Caroline

Publication Date
2018

Peer reviewed|Thesis/dissertation

Integrating Genetics Into Population-Based Studies

by

Caroline G. Tai

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Epidemiology and Translational Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Dedication

I would first and foremost like to thank my mentor and Dissertation Chair, Dr. John Witte for all the support and guidance he has given me over the years. You took a chance and gave me an opportunity to work on a research area that I was extremely excited about but that I may have underestimated. I was initially overwhelmed by the amount of knowledge and list of skills required of me. But under your patience and mentorship, I feel emboldened, ready to tackle the world's newest problems and unafraid of working in topic areas that constantly challenge me. Not only have you shown me how to be a confident scientist, but also how to be a balanced person, to pursue my interests both in work and in life. Thank you for everything you have taught me.

It is certain that I would not have made it through my dissertation without my Dissertation Committee member, Dr. Thomas Hoffmann. Thank you for always answering my endless questions and sharing all your lessons learned from your experiences. And you always do it with a smile! Thank you for your constant warmth and support. Thank you to Dr. Elad Ziv, my Dissertation Committee member, who always asked the most insightful questions and made me think critically about what I was doing and what the best approach should be. Your wisdom and clarity have been like a guiding light. And to my Dissertation Committee member, Dr. June Chan, you have been my role model as a female scientist and I aspire to be able to emulate how you are able to gather people around you and lead by example. Thank you for showing me how it can be done.

I will always be thankful for the patience and open-mindedness that my Epidemiology and Translational Science PhD Program Director, Dr. Maria Glymour has shown me. Thank you for taking the time to talk me through all my concerns over the years. And for all those times we

showed up in your office as a ball of frayed nerves, your reassuring words and advice were much appreciated. A big thank you to Dr. Robert Hiatt for taking me under your wing when I first arrived to UCSF. I enjoyed working with you and hearing about the inner workings of cancer registries and cancer centers through your unique perspective, I am truly grateful to have had the opportunity to learn from you.

Next, I thank my cohort-mates, Tu My To and Sarah Ackley for making my time at UCSF not just about work and research but also about camaraderie and finding a support network that's willing to listen to any crazy thing you think you are dealing with. Thank you for keeping me grounded and sane. Tu My, I will always remember our post-quals road trip fondly. And Sarah, we will always have Star Trek, live long and prosper my friends. Additionally, to all my fellow program-mates, I enjoyed learning about other areas of research through you. The incredible research diversity in our Epidemiology and Translational Science PhD program is astounding and exposed me to many research questions and analytical challenges that I would have never thought of. Thank you all for opening my mind. I would like to give a thank you to all the various groups, organizations, and people I've worked with at UCSF on issues of mentorship training, diversity, and improving the trainee experience at UCSF. Every time I meet faculty, staff, and students who contribute incredible amounts of their own personal free time to work on projects that benefit the community, it restores my faith in the altruistic spirit of my peers.

I owe my eternal gratitude to my lab-mates; the Witte Lab has been the most pleasant surprise of my entire PhD experience. Whether I need someone to bounce ideas off of, someone to turn to for programming questions, a sanity check on a statistical concept, a conference buddy,

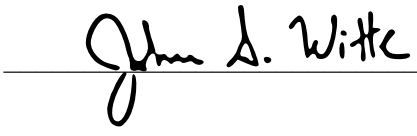
or a restaurant recommendation, my lab-mates have been there for me through everything. I am blessed to have the support of these collaborators of both work and fun.

I am most grateful for my family. My parents, who have always pushed me to do more, but would always support me if I felt stretched too far. Thank you for the ups and the downs, the laughs and the tears, it has all made me who I am today. And thank you to Roger Lee for the food and shelter, and for putting up with all my late night work-a-thon sessions. Your support and understanding have been invaluable to me in reaching this final stretch.

Finally, I would like to thank all the research participants whose data were essential in allowing me to conduct my dissertation work. Without you, none of our work is possible and I promise to continue being a responsible data steward of your research data.

Acknowledgments

A version of Chapter 2 of this dissertation has been published in *Birth Defects Research Part A: Clinical and Molecular Teratology*[1]. The published material is substantially the product of Caroline G. Tai's period of study at UCSF and was primarily conducted and written by her. The work she completed for this published manuscript is comparable to a standard dissertation chapter.

Approved:  John S. Witte, PhD, Dissertation Chair

[1] Tai CG, Graff RE, Liu J, Passarelli MN, Mefford JA, Shaw GM, et al. Detecting gene-environment interactions in human birth defects: Study designs and statistical methods: GxE Interaction Approaches for Birth Defects. *Birt Defects Res A Clin Mol Teratol.* 2015 Aug;103(8):692–702.

Integrating Genetics into Population-Based Studies

By

Caroline G. Tai

Abstract

As genotyping has become more affordable, there has been a growing effort to leverage genetic data to improve our understanding of the determinants of disease at the population level. New challenges arise with more widespread use of genetic data in epidemiological studies, from developing statistical methods and study designs that leverage genetic data to answering new bioethical questions around sharing research data with genetic information.

This dissertation was comprised of three projects that showcase important topics for epidemiologists to consider when using genetic data. These include developing genetic risk scores, determining the analysis method most suitable for the available data; and understanding the ethical, legal and social implications of widely sharing biobank data, particularly genetic data.

The first chapter explores differential genetic susceptibility between aggressive and non-aggressive prostate cancer and whether these differences can be used to develop a polygenic risk score for disease aggressiveness. This study was conducted in a Kaiser Permanente patient population electronic medical record and genotype data. Results from this analysis identified a subset of the known prostate cancer variants that are able to differentiate between aggressive and non-aggressive disease. The second chapter contrasts different gene-environment interaction parameters from various study designs using birth defects as the case example. This review aimed to clarify the options for conducting gene-environment interaction studies in the National

Birth Defects Prevention Study. Lastly, the third chapter utilizes qualitative methods to explore the perceived harms and benefits of sharing biobank data, which includes genetic information, from the perspective of several stakeholder groups involved in the development of a Kaiser Permanente biobank, the Research Program on Genes, Environment, and Health. Results from this investigation revealed that although the benefits of data sharing are clear, it is paramount that biobank participants are given the opportunity to be involved in the data governance and/or access process.

The lessons learned from this dissertation provide evidence that genetic data allows for exploration of inherited susceptibilities that can influence risk of disease but care must be taken to understand which groups of individuals should be compared and what questions are being answered, both scientifically and ethically.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

List of Figures

Chapter 1

Chapter 2

Chapter 1

Elucidating Genetic Differences between Aggressive and Non-Aggressive Prostate Cancer: Development of an Aggressive Disease Specific Polygenic Risk Score

Introduction

In the United States, prostate cancer (PrCa) is the most commonly diagnosed non-skin cancer among males, with 164,690 new cases expected in 2018 (1). While the majority of PrCa cases are indolent, the aggressive cases render PrCa the second most common cause of cancer death among men (1). These distinct disease courses underline the considerable heterogeneity of the disease. For example, when subdivided by stage, the 5-year survival for those diagnosed with distant disease is only 30%, while it is 100% for those with local or regional disease at diagnosis (1). Ideally, men with less aggressive disease would avoid treatments that 1) are expensive, 2) have adverse side effects, and 3) are unlikely to improve what is already a high survival rate. In contrast, those with aggressive disease would receive more intensive treatments. Encouraging men to choose less invasive treatment options relies heavily on the ability to accurately predict disease aggressiveness at the time of diagnosis. While clinical and pathologic measures help stratify patients to some degree, there remains room for improvement.

Previous studies have shown that most germline variants associated with overall PrCa are not associated with disease (2). A meta-analysis of four large genome-wide association studies (GWAS) only yielded two single nucleotide polymorphisms (SNPs) associated with aggressive disease and only one was replicated in another dataset (3). Additionally, many suspect that some PrCa variants may be associated with PSA and not with aggressive disease (4). We aimed to investigate whether any variants known to be associated with overall PrCa are only associated

with aggressive disease and to identify any germline genetic differences underlying aggressive and non-aggressive disease development.

Methods

Study Population and Defining PrCa Aggressiveness

Participants were included from three Kaiser Permanente (KP) case-control studies: the Research Program on Genes, Environment and Health (RPGEH), the ProHealth Study, and the California Men's Health Study (5). Controls were recruited between 2008-2012. We restricted our primary analyses to non-Hispanic white men, as determined by self-reported race / ethnicity, with confirmation by genetic principal components analysis. We replicated our results in African-American men. We included only unrelated men by filtering out 1st and 2nd degree relatives (`-unrelated` option, KING software v2.0) (6). Sample sizes for Hispanic and Asian populations were too limited and thus not included in these analyses.

PrCa status was determined from the KP California Cancer Registries or review of electronic health records available from 1994 to 2015 (5). For the non-Hispanic Whites there were 6,132 men with PrCa (cases) and 27,178 controls, and for the African Americans there were 664 cases and 1,719 controls.

For all cases, we abstracted information on Gleason grade, age at PrCa diagnosis, and all available PSA values measured during routine care, these included PSA values from PSA screening pre-diagnosis, diagnostic PSA, and post-treatment. From these data, we defined aggressive PrCa cases as any one of the following: 1) high Gleason grade ($\geq$ 7, with 4+3 as high); 2) early-onset disease (< 55 years of age); or 3) biochemical recurrence (BCR). BCR was defined as a PSA $\geq$ 0.4 ng/mL following surgery, or a PSA $\geq$ 2 ng/mL above the nadir following radiation therapy (7,8). Note that among individuals with Gleason Score of 7 (n=1839), which

can represent sub-scores of 3+4 or 4+3, 406 non-Hispanic white and 36 African American PrCa cases were missing sub-scores and were conservatively assumed to be 3+4. Previous studies have shown that scores of 3+4 and 4+3 are not equivalent and thus we decided to classify 3+4 as low Gleason grade and 4+3 as high Gleason grade (9,10). Cases without any of the aforementioned disease characteristics were defined as non-aggressive. For the non-Hispanic Whites there were 1,807 aggressive and 4,325 non-aggressive cases. For the African Americans there were 226 aggressive and 438 non-aggressive cases.

The KP Northern California Institutional Review Board and the University of California, San Francisco Human Research Protection Program Committee on Human Research approved this study.

Genotyping and Imputation

All study subjects were genotyped using ancestry-specific Affymetrix Axiom arrays (11,12), which included over 650,000 single nucleotide polymorphisms (SNPs). Two Axiom reagent kits were used in the genotyping process. To control for potential batch effects, we excluded SNPs that were associated with kit version ($p < 0.05$). Further quality control steps entailed excluding genotyped SNPs that had a minor allele frequency (MAF) < 0.01, a call rate < 95%, or that were out of Hardy-Weinberg equilibrium ($p < 1x10^{-5}$) in ancestry-specific control groups. We imputed unmeasured SNPs based on reference genomes from the 1000 Genomes Project using Impute2 v2.3.1 (13–15), yielding over 10 million SNPs. Following imputation, quality control filters were again applied, filtering out SNPs with information quality <0.5 and MAF < 0.05. There remained 7,115,637 SNPs for our analyses. All genomic locations reported are based on Genome Reference Consortium Human Build 37 (GRCh37).

Statistical Analyses

GWAS

We first searched genome-wide for SNPs associated with disease aggressiveness. The following three phenotype contrasts were analyzed with logistic regression: 1) aggressive PrCa cases versus controls ("aggressive case-control"); 2) non-aggressive PrCa cases versus controls ("non-aggressive case-control"); and 3) aggressive PrCa cases versus non-aggressive PrCa cases ("case-only"). GWAS analyses were adjusted for the following covariates: age at diagnosis (cases) or enrollment (controls), age^2 (to account for non-linear effects of age), kit, and the top 10 principal components of ancestry.

Polygenic Risk Scores

We further evaluated how the 167 known PrCa risk SNPs are associated with aggressiveness using a polygenic risk score (PRS) in logistic regression models for each of the three comparisons (aggressive and non-aggressive case-control and case-only). A complete list of the 167 known risk SNPs and their imputation information quality in our dataset are given in **Supplemental Table 1**. For each of the SNPs, we oriented the alleles (risk versus referent) such that the per-allele effect estimate was in the risk increasing direction. To construct the PRS, we calculated a weighted sum of the risk allele counts $PRS = \sum_k \hat{\beta}_k x_k$ where k is the number of SNPs included in the PRS, $\hat{\beta}_k$ is the natural log odds ratio of SNP k's association with prostate cancer, and x_k is the risk allele count for SNP k. To construct a PRS that represents the known PrCa risk SNPs, hereafter referred to as PRS_{167}, we used previously published log odds ratios for the 167 risk SNPs' associations with overall prostate cancer, which were mostly from studies of European ancestry individuals independent from this cohort (16,17). Odds ratios for the PRS_{167} calculation are provided in **Supplemental Table 1.**

Next, we evaluated whether any of the 167 known PrCa risk SNPs were differentially associated with aggressive versus non-aggressive disease. We aimed to investigate whether a PRS including only SNPs exhibiting statistically significant associations in the aggressive or non-aggressive case-control analyses could differentiate between aggressive and non-aggressive PrCa in a case-only analysis. Doing so entailed constructing two new PRS's: 1) PRS_A comprised of risk SNPs significant in the aggressive case-control analysis; and 2) PRS_{NA} comprised of risk SNPs significant in the non-aggressive case-control analysis. For the referenced case-control analyses, we randomly split the controls into two equal sized groups and used one group each for the aggressive and non-aggressive analyses. Statistical significance for SNP selection was based on a Bonferroni correction for the number of comparisons in each model (i.e., 0.05/167 SNPs = 0.0003). SNPs significant in both case-control analyses were excluded from PRS_A and PRS_{NA} because they would be unlikely to differentiate between the aggressive and non-aggressive phenotypes in a case-only analysis. Additionally we constructed a PRS combining the SNPs in PRS_A and PRS_{NA} creating PRS_{A+NA}, which may be even more strongly associated with aggressive disease than either PRS alone. The weights for these PRS's (i.e., log odds ratios) were from their respective case-control analyses in our data. In constructing the $PRS_{A+NA,}$ the odds ratios of the PRS_{NA} SNPs were reversed by multiplying the log odds ratio by -1 since we expected those SNPs to show a protective effect in a case-only analysis. The PRS_{NA} SNPs needed to be re-oriented in the risk increasing direction for aggressive disease to match the PRS_A SNPs, which were not modified. All PRS's were scaled to have a mean of zero and standard deviation of one, therefore the interpretation for all odds ratios reported is the increase in odds of disease for one standard deviation from the mean in the PRS.

Once PRS_A and PRS_{NA} were developed, we evaluated their association with diagnosis of aggressive PrCa in a case-only analysis. We acknowledge that a PRS constructed using odds ratios from the same study sample used to evaluate performance of the PRS, the non-Hispanic Whites in this study, can be biased and give an over-estimation of the strength of association. Therefore, we used the African Americans as a replication sample. Typically, a replication sample would be of the same genetic ancestry, i.e., individuals of non-Hispanic European ancestry. However replication in African Americans could indicate that the developed PRS's represent aggressive prostate cancer genes and pathways that are applicable across ancestry populations, even if SNPs are not perfectly tagging the true causal SNPs.

Additionally, we evaluated whether any case-only associations with the PRS's were primarily driven by one of the three aggressiveness measures. We did so by analyzing the associations between each of the PRS's (PRS_{167}, PRS_A, PRS_{NA}, PRS_{A+NA}) and: 1) High Gleason score; 2) early age of onset; and 3) BCR.

All of the analyses were initially undertaken in non-Hispanic Whites, since they comprised the largest proportion of our study subjects. We then attempted to replicate our findings in the African-American study subjects. All analyses adjusted for age at enrollment into the study (controls) or age at diagnosis (cases), Axiom array kit, and the top 10 principal components (PLINK v1.9), with exception to the early-onset analyses, which were not adjusted for age.

Pathway Associations

To evaluate whether the involvement of PRS_A SNPs in biological pathways are distinct from those regulated by the PRS_{NA} SNPs, we mapped each set of SNPs to a set of genes' encoding transcription factors based on the co-localization of the SNPs with each gene's

corresponding transcription factor binding sites (TFBS) via RegulomeDB v1.1 (18). We also included variants in high linkage disequilibrium (LD $r^2 > 0.8$) with each PRS_A or PRS_{NA} SNP, as identified by the LDproxy tool in LDlink (19), using the EUR super-population for LD structure.

Using these two lists of genes' transcription factors, for which TFBS co-localized with the SNPs in our two PRS's, we evaluated whether particular biological pathways were over- or under-represented by either gene list across the Reactome pathway database (Reactome version 58) (20,21). This was implemented using the pantherdb.org gene list analysis tool via the Statistical overrepresentation test, which uses a Fisher's exact test with a multiple testing corrected False Discovery Rate (22–25).

Results

Among the aggressive cases, the respective proportions with high Gleason score (≥ 4+3), with early-onset disease, or that experienced BCR were 38.7%, 15.4%, and 57.3% in non-Hispanic Whites, and 33.2%, 32.7%, and 51.3% in African Americans. Note that some individuals experienced more than one aggressive phenotype, and thus these percentages do not sum to 100% (**Table 1**). Among those with BCR, the median time to recurrence was 4.8 years in non-Hispanic Whites and 5.2 years in African Americans.

GWAS

The GWAS confirmed previously discovered loci at 8q24 and 11q13 (genome-wide significance $P < 1x10^{-8}$) in both aggressive and non-aggressive case-control analyses. Previously reported signals at 12q13 and 17q24 reached genome-wide significance in the aggressive case-control analysis but not in the non-aggressive case-control analysis. Conversely, previously reported signals at 10q11, 17q12, and 19q13 reached genome-wide significance in the non-

aggressive case-control analysis only (**Figure 1 & 2**). These results indicate that some previously identified loci may only be associated with aggressive or non-aggressive PrCa.

In the case-only GWAS, only the 19q13 locus was statistically significant after genome-wide Bonferroni correction. This locus contains *KLK3*, the gene that encodes prostate specific antigen (PSA) (**Figure 3**). One of the previously reported PrCa SNPs in this region known to affect PSA levels, rs2735839, was associated with an increased PSA at diagnosis among non-aggressive cases but not among aggressive cases in the non-Hispanic Whites ($p=0.008$ versus $p=0.33$, respectively).

Polygenic Risk Scores

Among the 167 known PrCa SNPs evaluated, 15 SNPs were statistically significantly associated with aggressive disease and 21 SNPs were associated with non-aggressive disease ($p<0.0003$). Between these two sets of SNPs, 10 were overlapping (**Figure 4**). There were thus five SNPs uniquely associated with aggressive disease and 11 uniquely associated with non-aggressive disease (**Figure 4**). The case-control results for these three sets of SNPs are summarized in **Table 2**.

In our PRS analyses, we first considered all 167 known PrCa SNPs. This PRS_{167} was not associated with a diagnosis of aggressive disease in case-only analyses of non-Hispanic Whites ($p=0.78$) or African Americans ($p=0.85$) (**Table 3**).

Next, we focused on the subsets of the 167 SNPs associated exclusively with aggressive disease, and, separately, non-aggressive disease. PRS_A included the five SNPs statistically significantly associated with aggressive disease and PRS_{NA} included the 11 SNPs statistically significantly associated with non-aggressive disease (**Figure 4 & Table 2**). We did not include

the 10 SNPs associated with both aggressive and non-aggressive disease in PRS_A or PRS_{NA} because they would be unlikely to differentiate between the phenotypes in a case-only analysis.

PRS_A was strongly associated with a diagnosis of aggressive disease in the case-only analysis of non-Hispanic Whites ($p=1.4x10^{-6}$). This finding was replicated in the independent set of African-American samples (p=0.007) (**Table 3**). Focusing on the components of the aggressiveness phenotype, PRS_A was associated with Gleason score (p=0.007) and BCR ($p=1.7x10^{-4}$) in the non-Hispanic White men. However, these associations were much weaker in the African Americans (Gleason p=0.09, BCR p=0.08) (**Table 3**). PRS_A was not associated with early age of onset in non-Hispanic Whites (p=0.10) or in African Americans (p=0.89). We also evaluated whether the PRS_A could differentiate aggressive from non-aggressive disease in a population that also includes controls without PrCa. This analysis showed that the PRS_{167} showed strong association with both aggressive ($p=6.2x10^{-104}$ in Non-Hispanic Whites, $p=4.0x10^{-8}$ in African Americans) and non-aggressive disease in case-control analyses ($p=3.6x10^{-197}$in non-Hispanic Whites, $p=2.5x10^{-9}$ in African Americans). But the PRS_A was only statistically significant for aggressive disease (p=0.008) and not with non-aggressive disease in the African Americans (p=0.636) in the case-control analyses (**Table 4**).

PRS_{NA} was inversely associated with a diagnosis of aggressive disease in the case-only analysis of non-Hispanic Whites (OR=0.89, 95%CI: 0.84-0.94, $p=4.0 \times 10^{-5}$). This observation was not replicated in the corresponding analysis of African Americans (OR=1.03, 95%CI: 0.85-1.23, p=0.78) (**Table 3**).

Finally, a new PRS that combined the five PRS_A and 11 PRS_{NA} SNPs was not associated with a diagnosis of aggressive disease in the case-only analysis, thus results were not included here.

Pathway Associations

There were some shared Reactome pathways associated with both the PRS_A and the PRS_{NA} SNPs, such as Developmental Biology (FDR = $4.2x10^{-8}$ and $4.2x10^{-10}$, respectively), Activation of HOX genes during differentiation (FDR = $1.1x10^{-8}$ and $1.1x10^{-7}$, respectively), and Transcriptional Regulation by TP53 (FDR = $3.4x10^{-6}$, $2.5x10^{-5}$, respectively). But some pathways were uniquely associated with PRS_A or PRS_{NA} SNPs. For example, chromatin organization was associated with PRS_A SNPs but not associated with PRS_{NA} SNPs (FDR = $4.7x10^{-8}$ and 0.306, respectively) while myogenesis was associated with PRS_{NA} SNPs but not with PRS_A SNPs (FDR = 0.007 and 0.413, respectively). Results for all pathways statistically significant with aggressive or non-aggressive disease are included in **Table 5**.

Sensitivity Analyses

False Discovery Rate (FDR) is an alternative to using Bonferroni corrected significance thresholds in the SNP selection procedure. To explore whether the developed PRS's are greatly influenced by the significance threshold used, we repeated the analysis using FDR < 0.05 for selecting the SNPs that would be included in the aggressive and non-aggressive risk scores. The new FDR based risk scores included more SNPs than the Bonferroni correction based risk scores. Ten SNPs were eligible for inclusion in the aggressive FDR based risk score ($PRS_{FDR,A}$) and 30 SNPs in the non-aggressive FDR based risk score ($PRS_{FDR,NA}$). The associations of the $PRS_{FDR,A}$ with aggressive disease were weaker compared to the results from the primary analysis using Bonferroni correction, p=$1.7x10^{-4}$ in non-Hispanic Whites and p=0.052 in African Americans. Conversely, the associations of the $PRS_{FDR,NA}$ was slightly stronger though still not

statistically significant in the African Americans, $p=7.60x10^{-7}$ in non-Hispanic Whites and p=0.581 in African Americans (**Table 6**). This may indicate that choice of significance threshold may affect these results but this does not seem to change to the main implication that only a subset of the known PrCa loci are associated with aggressive disease.

In defining BCR, it is common to require a minimum follow-up period among cases, for example, at least three years to ensure that sufficient time has passed to observe a recurrence. In our data, 358 PrCa cases among the non-Hispanic Whites and 45 PrCa cases among the African Americans had less than 3 years follow-up time and classified as not experiencing BCR. To evaluate the impact of these individuals on our results, we re-estimated the association of the PRS's with the outcomes composite aggressive disease and BCR, while excluding these individuals (**Table 7**). We did not find the results to drastically differ. The analyses involving early-onset and high Gleason would be unaffected and thus were not re-estimated.

Discussion

We found that five known PrCa loci may be associated with aggressive PrCa but not with non-aggressive PrCa. In addition, we determined that a PRS constructed based on these five SNPs predicted diagnosis with aggressive disease better than a PRS including all known PrCa risk loci. This highlights the importance of SNP selection for score construction based on relevant phenotype associations rather than maximal number of SNPs.

In our genome-wide analysis data, only the 19q13 region yielded variants that were differentiated between aggressive and non-aggressive PrCa. Others have also found that the 19q13 SNP, rs2735839, is associated with aggressive disease in case-only frameworks. These results, however, may largely be a artifact of screening bias since this observation may reflect constitutive PSA levels and the association arising due to PSA screening for PrCa (26–28).

Some associations we found between our PRS_A and PRS_{NA} SNPs and aggressive disease have been identified in previous studies. Among the five SNPs that comprised PRS_A, rs7968403 is located in the intronic region of *RASSF3*, which is a member of the Ras association domain family protein (RASSF) gene family and previously implicated as a potential tumor suppressor gene (29). Among the eleven SNPs that comprised the non-aggressive PRS, rs2735839 has been previously reported to be an aggressiveness SNP (30). However, others have noted that the risk increasing allele seems to be associated with less aggressive disease, which is consistent with our results (31). The SNPs, rs9364554 and rs10486567 have also been shown to be associated with overall prostate cancer but not specifically with aggressiveness, again consistent with our results (32). In the group of SNPs associated with both aggressive and non-aggressive PrCa, rs10993994 was previously reported to be associated with BCR (33).

We identified some Reactome pathways associated with both aggressive and non-aggressive SNPs but also pathways that were differentially associated with each set of SNPs, for example, chromatin organization with aggressive SNPs and myogenesis with non-aggressive SNPs. This could imply that the etiology and underlying biological pathways involved in disease development are differently between aggressive and non-aggressive disease.

Although we were able to replicate the aggressive PRS in African-American individuals, differences in LD structure between genetic ancestral groups make it tenuous to apply a PRS constructed based on European ancestry populations to an African ancestry population. Our identification of a significant association in this study indicates that a PRS constructed of SNPs discovered in and weighted by effect estimates calculated in African ancestry populations would likely show a stronger effect. Additionally, because clinical data were from passive collection during routine medical care, there were some missing data (i.e., Gleason score, PSA levels).

Though missing data were likely to be missing completely at random, this assumption cannot be verified. This study was limited in sample size, and may have been underpowered to detect some associations. It is likely that additional aggressive specific SNPs exist, even among the known loci, but we may have been underpowered to detect them. Identification of additional aggressiveness SNPs could improve the performance of the PRS_A we developed.

Despite these limitations, this study is one of the first comprehensive evaluations to determine whether known PrCa loci are specific to aggressive or non-aggressive disease. Our approach leveraging both case-control and case-only comparisons provides better clarity in the search for SNPs associated with aggressive PrCa. As case-only approaches are likely to identify PSA-related associations, comparisons of cases to true controls without cancer may reveal other biological mechanisms associated with aggressive PrCa unrelated to genetic contributions to PSA levels. Aggressive specific PRS can provide information orthogonal to PSA screening results that could help clinicians and patients to determine appropriate monitoring and treatment decisions for prostate cancer.

References

1. American Cancer Society. Cancer Facts & Figures 2018. Atlanta: American Cancer Society; 2018.

2. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, et al. Cumulative Association of Five Genetic Variants with Prostate Cancer. N Engl J Med. 2008 Feb 28;358(9):910–9.

3. Amin Al Olama A, Kote-Jarai Z, Schumacher FR, Wiklund F, Berndt SI, Benlloch S, et al. A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. Hum Mol Genet. 2013 Jan 15;22(2):408–15.

4. Penney KL, Schumacher FR, Kraft P, Mucci LA, Sesso HD, Ma J, et al. Association of KLK3 (PSA) genetic variants with prostate cancer risk and PSA levels. Carcinogenesis. 2011 Jun;32(6):853–9.

5. Hoffmann TJ, Van Den Eeden SK, Sakoda LC, Jorgenson E, Habel LA, Graff RE, et al. A Large Multiethnic Genome-Wide Association Study of Prostate Cancer Identifies Novel Risk Variants and Substantial Ethnic Differences. Cancer Discov. 2015 Aug 1;5(8):878–91.

6. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73.

7. Roach M, Hanks G, Thames H, Schellhammer P, Shipley WU, Sokol GH, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: Recommendations of the RTOG-ASTRO Phoenix Consensus Conference. Int J Radiat Oncol. 2006 Jul;65(4):965–74.

8. Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Lilja H, et al. Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. J Clin Oncol Off J Am Soc Clin Oncol. 2006 Aug 20;24(24):3973–8.

9. Stark JR, Perner S, Stampfer MJ, Sinnott JA, Finn S, Eisenstein AS, et al. Gleason Score and Lethal Prostate Cancer: Does 3 + 4 = 4 + 3? J Clin Oncol. 2009 Jul 20;27(21):3459–64.

10. Wright JL, Salinas CA, Lin DW, Kolb S, Koopmeiners J, Feng Z, et al. Prostate cancer specific mortality and Gleason 7 disease differences in prostate cancer outcomes between

cases with Gleason 4 + 3 and Gleason 3 + 4 tumors in a population based cohort. J Urol. 2009 Dec;182(6):2702–7.

11. Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics. 2011 Dec;98(6):422–30.

12. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, et al. Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. Genomics. 2011 Aug;98(2):79–89.

13. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012 Jul 22;44(8):955–9.

14. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 Bethesda Md. 2011 Nov;1(6):457–70.

15. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009 Jun;5(6):e1000529.

16. Han Y, Signorello LB, Strom SS, Kittles RA, Rybicki BA, Stanford JL, et al. Generalizability of established prostate cancer risk variants in men of African ancestry: Generalizability of prostate cancer SNPs in African ancestry. Int J Cancer. 2015 Mar 1;136(5):1210–7.

17. Schumacher FR. Prostate cancer meta-analysis of more than 140,000 men identifies 63 novel prostate cancer susceptibility loci. Nat Genet. In Press;X(X):XX–XX.

18. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012 Sep;22(9):1790–7.

19. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinforma Oxf Engl. 2015 Nov 1;31(21):3555–7.

20. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018 Jan 4;46(D1):D649–55.

21. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, et al. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers. 2012 Nov 8;4(4):1180–211.

22. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017 Jan 4;45(D1):D183–9.

23. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013 Jul 18;8(8):1551–66.

24. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res. 2006 Jul 1;34(Web Server):W645–50.

25. Mi H, Thomas P. PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools. In: Nikolsky Y, Bryant J, editors. Protein Networks and Pathway Analysis [Internet]. Totowa, NJ: Humana Press; 2009 [cited 2018 May 9]. p. 123–40. Available from: http://link.springer.com/10.1007/978-1-60761-175-2_7

26. Helfand BT, Roehl KA, Cooper PR, McGuire BB, Fitzgerald LM, Cancel-Tassin G, et al. Associations of prostate cancer risk variants with disease aggressiveness: results of the NCI-SPORE Genetics Working Group analysis of 18,343 cases. Hum Genet. 2015 Apr;134(4):439–50.

27. Ahn J, Berndt SI, Wacholder S, Kraft P, Kibel AS, Yeager M, et al. Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. Nat Genet. 2008 Sep;40(9):1032–4.

28. Bensen JT, Xu Z, Smith GJ, Mohler JL, Fontham ETH, Taylor JA. Genetic polymorphism and prostate cancer aggressiveness: A case-only study of 1,536 GWAS and candidate SNPs in African-Americans and European-Americans. The Prostate. 2013 Jan;73(1):11–22.

29. Tommasi S, Dammann R, Jin S-G, Zhang X, Avruch J, Pfeifer GP. RASSF3 and NORE1: identification and cloning of two human homologues of the putative tumor suppressor gene RASSF1. Oncogene. 2002 Apr;21(17):2713–20.

30. He Y, Gu J, Strom S, Logothetis CJ, Kim J, Wu X. The Prostate Cancer Susceptibility Variant rs2735839 Near KLK3 Gene Is Associated with Aggressive Prostate Cancer and Can Stratify Gleason Score 7 Patients. Clin Cancer Res. 2014 Oct 1;20(19):5133–9.

31. Kader AK, Sun J, Isaacs SD, Wiley KE, Yan G, Kim S-T, et al. Individual and cumulative effect of prostate cancer risk-associated variants on clinicopathologic variables in 5,895 prostate cancer patients. The Prostate. 2009 Aug 1;69(11):1195–205.

32. International Consortium for Prostate Cancer Genetics, Teerlink CC, Thibodeau SN, McDonnell SK, Schaid DJ, Rinckleb A, et al. Association analysis of 9,560 prostate cancer cases from the International Consortium of Prostate Cancer Genetics confirms the role of reported prostate cancer associated SNPs for familial disease. Hum Genet. 2014 Mar;133(3):347–56.

33. Huang S-P, Huang L-C, Ting W-C, Chen L-M, Chang T-Y, Lu T-L, et al. Prognostic significance of prostate cancer susceptibility variants on prostate-specific antigen recurrence after radical prostatectomy. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2009 Nov;18(11):3068–74.

Tables & Figures

Table 1.1. Participant characteristics

Characteristics	Non-Hispanic Whites			African Americans		
	Controls, n= 27178	Non-aggressive Cases, n= 4325	Aggressive Cases, n= 1807	Controls, n= 1719	Non-aggressive Cases, n= 436	Aggressive Cases, n= 226
Age at diagnosis or study enrollment [a], %						
≤ 50	10.0	0	2.9	8.0	0.0	8.8
50-60	18.5	11.8	20.8	22.0	19.2	38.9
60-70	32.9	49.5	37.9	37.3	55.5	35.0
70-80	26.2	31.7	31.1	27.7	22.4	15.9
>80	12.3	7.0	7.4	4.9	3.0	1.3
BMI, median (IQR)	26 (24 - 29)	26 (24 - 29)	26.43 (24 - 29)	28 (25 - 31)	28 (25.40 - 31)	28.69 (26 - 32)
Experienced Biochemical Recurrence, %	NA	NA	57.3	NA	NA	51.3
Early Onset Diagnosis, %	NA	NA	15.4	NA	NA	32.7
Gleason Score [b], %						
2-6	NA	72.9	33.2	NA	66.9	41.1
7 (3+4)		27.2	17.4		33.1	21.1
7(4+3)		0	18.9		0	17.4
8-10		0	30.5		0	20.5

[a] Age of study enrollment for controls and age of diagnosis for cases

[b] Some cases were missing Gleason information, 32% of non-Hispanic white PrCa cases and 25% of African American PrCa cases

NA = Not Applicable

Table 1.2. Associations between single nucleotide polymorphisms and aggressive or non-aggressive prostate cancer

Region	SNP	Risk Allele	Ref Allele	RAF	Odds Ratio	95% CI	P-value
Aggressive SNPs [a]							
1q21	rs17599629	G	A	0.21	1.19	(1.09 , 1.30)	4.78E-05
2p11	rs10187424	A	G	0.58	1.16	(1.08 , 1.25)	3.89E-05
2q31	rs12621278	A	G	0.94	1.43	(1.20 , 1.69)	4.77E-05
12q13	rs902774	A	G	0.14	1.30	(1.19 , 1.43)	3.42E-08
12q14	rs7968403	T	C	0.64	1.16	(1.07 , 1.25)	1.62E-04
Non-Aggressive SNPs [a]							
1q21	rs34579442	C	CT	0.33	1.11	(1.05 , 1.17)	2.84E-04
3q13	rs7611694	A	C	0.60	1.11	(1.05 , 1.16)	1.49E-04
4q24	rs7679673	C	A	0.58	1.10	(1.04 , 1.16)	2.88E-04
6q25	rs9364554	T	C	0.27	1.14	(1.08 , 1.20)	3.85E-06
7p15	rs10486567	G	A	0.76	1.17	(1.10 , 1.25)	3.29E-07
8p21	rs1512268	A	G	0.43	1.18	(1.12 , 1.24)	2.32E-10
8q24	rs10086908	T	C	0.70	1.21	(1.15 , 1.29)	2.32E-11
8q24	rs445114	T	C	0.63	1.11	(1.05 , 1.16)	1.96E-04
11q25	rs878987	G	A	0.14	1.17	(1.09 , 1.25)	1.24E-05
12q13	rs10875943	C	T	0.29	1.13	(1.07 , 1.20)	9.69E-06
19q13	rs2735839	G	A	0.85	1.26	(1.16 , 1.35)	2.71E-09
Overlap SNPs [a]							
6q22	rs339331	T	C	0.70	1.14	(1.09 , 1.20)	2.46E-06
8q24	rs1016343	T	C	0.20	1.26	(1.20 , 1.33)	2.35E-17
8q24	rs16902104	T	C	0.13	1.22	(1.16 , 1.29)	3.03E-09
8q24	rs6983267	G	T	0.50	1.26	(1.21 , 1.32)	1.18E-23
8q24	rs11986220	A	T	0.09	1.55	(1.45 , 1.65)	3.83E-37
10q11	rs10993994	T	C	0.39	1.22	(1.17 , 1.28)	1.20E-17
11q13	rs10896449	G	A	0.51	1.20	(1.16 , 1.26)	5.21E-15
17q12	rs7501939	C	T	0.60	1.18	(1.13 , 1.23)	1.87E-11
17q24	rs1859962	G	T	0.48	1.16	(1.11 , 1.21)	1.21E-09
22q13	rs5759167	G	T	0.50	1.14	(1.09 , 1.19)	4.39E-07

Bonferroni corrected p-value threshold = 2.99E-04

RAF = Risk allele frequency in the non-Hispanic Whites

[a] Aggressive SNP effect estimates are from the aggressive case-control while Non-Aggressive SNP effect estimates are from the non-aggressive case control. For Overlap SNPs, odds ratio and P-values are based on meta-analysis of aggressive and non-aggressive case-control analyses

Table 1.3. Performance of PRS in Non-Hispanic Whites and African Americans across phenotypes in case-only analyses

Case-only Analysis	PRS_{167}			PRS_A			PRS_{NA}		
	Ratio (95% CI)		P-value	Ratio (95% CI)		P-value	Ratio (95% CI)		P-value
Non-Hispanic Whites									
Composite Aggressive	0.99	(0.94, 1.05)	0.778	1.15	(1.09, 1.22)	1.42E-06	0.89	(0.84,0.94)	4.01E-05
Early Onset	1.37	(1.22, 1.55)	3.86E-07	1.11	(0.98, 1.25)	0.098	0.99	(0.88,1.12)	0.892
High Gleason	0.90	(0.82, 0.99)	0.028	1.13	(1.03, 1.23)	0.007	0.80	(0.73,0.88)	9.26E-07
Biochemical Recurrence	1.02	(0.95, 1.09)	0.572	1.14	(1.06, 1.22)	1.69E-04	0.94	(0.88,1.01)	0.092
African Americans									
Composite Aggressive	1.02	(0.85, 1.22)	0.851	1.28	(1.07, 1.53)	0.007	1.03	(0.85,1.23)	0.784
Early Onset	1.19	(0.93, 1.53)	0.165	0.98	(0.76, 1.26)	0.893	1.10	(0.86,1.42)	0.444
High Gleason	1.04	(0.79, 1.36)	0.785	1.25	(0.96, 1.62)	0.091	0.97	(0.74,1.27)	0.833
Biochemical Recurrence	1.05	(0.85, 1.29)	0.646	1.20	(0.98, 1.47)	0.080	1.00	(0.81,1.24)	0.992

Table 1.4. Performance of PRS in Non-Hispanic Whites and African Americans across phenotypes in case-control analyses

Aggressive Case-control Analysis	PRS_{167}			PRS_A		
	Ratio (95% CI)		P-value	Ratio (95% CI)		P-value
Non-Hispanic Whites						
Composite Aggressive	1.73	(1.65, 1.82)	6.20E-104	1.27	(1.21, 1.33)	2.82E-22
Early Onset	2.05	(1.81, 2.32)	2.57E-29	1.26	(1.12, 1.42)	1.67E-04
High Gleason	1.54	(1.42, 1.67)	5.06E-26	1.24	(1.15, 1.35)	4.94E-08
Biochemical Recurrence	1.74	(1.63, 1.85)	1.24E-63	1.28	(1.21, 1.37)	5.35E-15
African Americans						
Composite Aggressive	1.53	(1.31, 1.78)	4.03E-08	1.21	(1.05, 1.40)	0.008
Early Onset	1.71	(1.26, 2.35)	0.001	1.13	(0.85, 1.51)	0.396
High Gleason	1.49	(1.15, 1.92)	0.002	1.30	(1.02, 1.64)	0.031
Biochemical Recurrence	1.56	(1.28, 1.91)	1.48E-05	1.23	(1.01, 1.49)	0.037
Non-aggressive Case-control Analysis	PRS_{167}			PRS_A		
	Ratio (95% CI)		P-value	Ratio (95% CI)		P-value
Non-Hispanic Whites						
Composite Non-Aggressive	1.70	(1.64, 1.76)	3.58E-197	1.11	(1.07, 1.14)	1.19E-09
Late/normal Onset	1.69	(1.64, 1.75)	9.46E-245	1.15	(1.11, 1.18)	1.57E-20
Low Gleason	1.75	(1.68, 1.82)	3.96E-184	1.12	(1.08, 1.16)	1.27E-09
No Biochemical Recurrence	1.70	(1.65, 1.76)	1.38E-229	1.13	(1.09, 1.16)	8.80E-15
African Americans						
Composite Non-Aggressive	1.41	(1.26, 1.58)	2.48E-09	1.03	(0.92, 1.15)	0.636
Late/normal Onset	1.42	(1.28, 1.57)	1.30E-11	1.09	(0.99, 1.20)	0.080
Low Gleason	1.46	(1.31, 1.64)	6.91E-11	1.05	(0.94, 1.17)	0.378
No Biochemical Recurrence	1.43	(1.29, 1.59)	9.81E-12	1.06	(0.96, 1.17)	0.262

Table 1.5. Pathway associations for aggressive and non-aggressive prostate cancer

Note: FDR is calculated by the Benjamini-Hochberg procedure, values are normally between 0 and 1 but can be greater than 1.0 due to multiple test correction. Values in the FDR columns are color-coded green > yellow > orange > red such that green indicates larger values while red indicates smaller values. Lists of genes used in the gene list analysis are provided below following the table.

	Aggressive SNPs Results			Non-aggressive SNPs Results		
Pathway	Fold Enrichment	Raw P value	FDR	Fold Enrichment	Raw P value	FDR
Regulation of TP53 Activity through Association with Co-factors	< 0.01	1.000	1.310	33.9	1.54E-04	0.014
Regulation of TP53 Activity through Acetylation	16.1	1.86E-04	5.44E-03	13.8	0.011	0.236
↳Regulation of TP53 Activity	4.9	3.41E-04	7.88E-03	9.4	8.26E-07	1.83E-04
↳Transcriptional Regulation by TP53	5.0	2.06E-08	3.41E-06	6.3	7.65E-08	2.54E-05
TFAP2 (AP-2) family regulates transcription of cell cycle factors	73.8	1.55E-06	1.55E-04	31.6	0.037	0.586
↳Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors	25.2	5.23E-10	1.30E-07	14.4	0.001	0.059
↳Generic Transcription Pathway	5.7	2.22E-23	4.43E-20	4.9	2.17E-11	2.16E-08
↳Gene Expression	3.3	2.15E-16	2.13E-13	3.0	1.69E-08	6.74E-06
TFAP2 (AP-2) family regulates transcription of other transcription factors	69.2	4.12E-05	0.002	< 0.01	1.000	1.230
TFAP2 (AP-2) family regulates transcription of growth factors and their receptors	36.9	1.18E-05	6.92E-04	15.8	0.067	0.818
TFAP2A acts as a transcriptional repressor during retinoic acid induced cell differentiation	36.9	0.002	0.039	< 0.01	1.000	1.280
Activation of the TFAP2 (AP-2) family of transcription factors	23.1	5.03E-04	0.010	26.4	0.003	0.107
Regulation of gene expression in beta cells	15.4	0.001	0.026	26.4	2.97E-04	0.018
Regulation of gene expression in early pancreatic precursor cells	39.6	1.38E-04	0.004	45.2	0.001	0.058
↳Regulation of beta-cell development	15.9	2.92E-05	0.001	32.7	7.66E-08	2.18E-05
POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation	61.5	0.001	0.021	52.7	0.025	0.454
↳Transcriptional regulation of pluripotent stem cells	16.3	4.02E-06	3.20E-04	18.6	9.30E-05	0.010
↳Developmental Biology	3.8	8.49E-11	4.22E-08	5.6	2.10E-13	4.17E-10
POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation	38.5	7.55E-07	9.39E-05	26.4	0.003	0.109
Nuclear Receptor transcription pathway	33.1	2.94E-16	1.95E-13	24.3	3.65E-07	9.08E-05
CDO in myogenesis	6.6	0.041	0.411	22.6	4.67E-05	0.007
↳Myogenesis	6.6	0.041	0.413	22.6	4.67E-05	0.007
MAPK targets/ Nuclear events mediated by MAP kinases	18.5	2.11E-06	1.91E-04	21.1	5.96E-05	0.007
↳MAP kinase activation in TLR cascade	10.8	7.47E-06	5.13E-04	13.2	5.53E-05	0.007
↳TRIF-mediated TLR3/TLR4 signaling	7.7	1.64E-05	8.83E-04	8.2	4.43E-04	0.024
↳MyD88-independent TLR3/TLR4 cascade	7.7	1.64E-05	8.37E-04	8.2	4.43E-04	0.023
↳Activated TLR4 signalling	6.7	4.10E-05	0.002	7.2	8.01E-04	0.039
↳Toll Like Receptor 4 (TLR4) Cascade	6.0	8.61E-05	0.003	6.4	0.001	0.057
↳Toll Like Receptor 3 (TLR3) Cascade	7.7	1.64E-05	8.59E-04	8.2	4.43E-04	0.023
↳Toll-Like Receptors Cascades	5.0	2.87E-04	0.007	5.3	0.003	0.102
↳MyD88 cascade initiated on plasma membrane	8.0	4.51E-05	0.002	9.8	2.10E-04	0.017
↳Toll Like Receptor 10 (TLR10) Cascade	8.0	4.51E-05	0.002	9.8	2.10E-04	0.016
↳Toll Like Receptor 5 (TLR5) Cascade	8.0	4.51E-05	0.002	9.8	2.10E-04	0.016
↳MyD88_Mal cascade initiated on plasma membrane	7.1	8.97E-05	0.003	8.7	3.51E-04	0.021
↳Toll Like Receptor TLR1_TLR2 Cascade	6.9	1.09E-04	0.004	8.4	4.04E-04	0.023
↳Toll Like Receptor 2 (TLR2) Cascade	6.9	1.09E-04	0.003	8.4	4.04E-04	0.022
↳Toll Like Receptor TLR6_TLR2 Cascade	7.1	8.97E-05	0.003	8.7	3.51E-04	0.021
↳TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	7.9	4.85E-05	0.002	9.7	2.22E-04	0.016
↳MyD88 dependent cascade initiated on endosome	7.7	5.59E-05	0.002	9.4	2.47E-04	0.016
↳Toll Like Receptor 9 (TLR9) Cascade	7.4	6.88E-05	0.002	9.1	2.88E-04	0.019
↳Toll Like Receptor 7/8 (TLR7/8) Cascade	7.7	5.59E-05	0.002	9.4	2.47E-04	0.017
Activation of anterior HOX genes in hindbrain development during early embryogenesis	13.0	5.51E-10	1.22E-07	20.5	2.21E-11	1.47E-08
↳Activation of HOX genes during differentiation	13.0	5.51E-10	1.10E-07	20.5	2.21E-11	1.10E-08
Nuclear Events (kinase and transcription factor activation)	11.5	0.003	0.048	19.8	6.35E-04	0.032
Transcriptional activation of mitochondrial biogenesis	13.8	9.30E-06	5.78E-04	19.8	9.05E-06	0.002
↳Mitochondrial biogenesis	11.3	2.66E-05	1.26E-03	16.1	2.24E-05	0.004

Pathway (continued)	Aggressive SNPs Results			Non-aggressive SNPs Results		
	Fold Enrichment	Raw P value	FDR	Fold Enrichment	Raw P value	FDR
Transcriptional regulation of pluripotent stem cells	16.3	4.02E-06	3.20E-04	18.6	9.30E-05	0.010
RNA Polymerase III Transcription Initiation From Type 3 Promoter	6.6	0.041	0.415	17.0	9.57E-04	0.044
↳RNA Polymerase III Transcription	6.8	0.012	0.150	15.4	1.82E-04	0.016
↳RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription	7.5	5.95E-06	4.23E-04	7.1	8.33E-04	0.040
YAP1- and WWTR1 (TAZ)-stimulated gene expression	19.1	1.78E-06	1.68E-04	16.4	0.001	0.048
RNA Polymerase III Abortive And Retractive Initiation	6.8	0.012	0.149	15.4	1.82E-04	0.015
BMAL1:CLOCK, NPAS2 activates circadian gene expression	13.2	1.20E-05	6.61E-04	11.3	0.003	0.099
↳Circadian Clock	10.4	9.10E-06	6.04E-04	12.8	6.41E-05	0.008
Transcriptional regulation of white adipocyte differentiation	9.5	3.99E-06	3.31E-04	12.2	1.49E-05	0.003
RNA polymerase II transcribes snRNA genes	9.2	1.89E-05	9.38E-04	11.3	1.10E-04	0.010
PPARA activates gene expression	9.2	7.66E-08	1.17E-05	8.6	9.24E-05	0.010
↳Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	9.9	9.54E-09	1.73E-06	8.5	1.02E-04	0.010
Binding of TCF/LEF:CTNNB1 to target gene promoters	39.6	1.38E-04	0.004	45.2	0.001	0.056
↳Formation of the beta-catenin:TCF transactivating complex	12.2	3.53E-06	3.06E-04	9.0	0.005	0.134
↳TCF dependent signaling in response to WNT	4.7	7.71E-05	0.003	3.3	0.037	0.598
↳Signaling by Wnt	3.5	4.16E-04	0.009	2.7	0.038	0.594
↳Signal Transduction	1.6	0.002	0.031	1.2	0.497	2.570
LRR FLII-interacting protein 1 (LRRFIP1) activates type I IFN production	36.9	0.002	0.039	31.6	0.037	0.591
G2 Phase	36.9	0.002	0.039	< 0.01	1.000	1.270
Nuclear Receptor transcription pathway	33.1	2.94E-16	1.95E-13	24.3	3.65E-07	9.08E-05
p75NTR negatively regulates cell cycle via SC1	30.8	0.003	0.050	< 0.01	1.000	1.180
Activation of the AP-1 family of transcription factors	27.7	3.21E-04	0.008	15.8	0.067	0.808
Repression of WNT target genes	21.3	6.14E-04	0.012	12.2	0.085	0.904
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	20.2	1.70E-07	2.42E-05	9.9	0.019	0.372
↳Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer	14.7	1.15E-06	1.20E-04	7.2	0.034	0.557
↳Signaling by TGF-beta Receptor Complex	9.0	2.23E-05	0.001	4.4	0.079	0.888
YAP1- and WWTR1 (TAZ)-stimulated gene expression	19.1	1.78E-06	1.68E-04	16.4	0.001	0.048
Signaling by cytosolic FGFR1 fusion mutants	17.3	0.001	0.020	9.9	0.102	1.020
↳FGFR1 mutant receptor activation	14.2	2.83E-04	0.007	6.1	0.157	1.350
↳Signaling by FGFR1 in disease	11.2	6.43E-04	0.013	4.8	0.193	1.450
↳Signaling by FGFR in disease	9.6	6.34E-05	0.002	5.5	0.055	0.721
↳Diseases of signal transduction	4.9	8.18E-07	9.57E-05	3.4	0.010	0.224
↳Disease	2.1	0.003	0.044	1.9	0.031	0.532
Polo-like kinase mediated events	17.3	0.001	0.020	19.8	0.006	0.141
Deactivation of the beta-catenin transactivating complex	15.4	8.69E-07	9.61E-05	7.5	0.031	0.535
G0 and Early G1	14.8	2.47E-04	0.006	< 0.01	1.000	4.440
↳Mitotic G1-G1/S phases	4.7	0.001	0.020	2.3	0.222	1.590
Activation of gene expression by SREBF (SREBP)	13.8	9.30E-06	5.97E-04	4.0	0.228	1.610
↳Regulation of cholesterol biosynthesis by SREBP (SREBF)	10.5	3.98E-05	0.002	3.0	0.289	1.850
Signaling by Retinoic Acid	13.2	1.20E-05	6.80E-04	11.3	0.003	0.101
Pre-NOTCH Transcription and Translation	12.7	4.12E-04	0.009	5.5	0.172	1.380
↳Pre-NOTCH Expression and Processing	8.2	0.002	0.033	3.5	0.252	1.710
↳Signaling by NOTCH	7.7	4.83E-06	3.70E-04	4.4	0.033	0.545
Oncogene Induced Senescence	12.3	4.63E-04	0.010	< 0.01	1.000	4.520
↳Cellular Senescence	5.4	6.84E-05	0.002	5.1	0.003	0.109
↳Cellular responses to stress	3.2	5.66E-04	0.012	3.2	0.007	0.183
HDMs demethylate histones	12.0	0.003	0.044	< 0.01	1.000	1.920
↳Chromatin modifying enzymes	7.4	1.42E-10	5.64E-08	3.5	0.015	0.309
↳Chromatin organization	7.4	1.42E-10	4.70E-08	3.5	0.015	0.306
TRAF6 mediated IRF7 activation	11.5	5.79E-04	0.012	4.9	0.188	1.430
↳RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	5.8	0.002	0.038	4.0	0.094	0.975
HDACs deacetylate histones	10.5	3.98E-05	0.002	3.0	0.289	1.840
NOTCH1 Intracellular Domain Regulates Transcription	10.3	1.94E-04	0.005	7.0	0.035	0.574
↳Signaling by NOTCH1	6.6	0.001	0.024	4.5	0.075	0.876
RNA Polymerase II Transcription Initiation	10.0	2.13E-04	0.006	10.3	0.004	0.109
↳RNA Polymerase II Transcription Initiation And Promoter Clearance	10.0	2.13E-04	0.006	10.3	0.004	0.104
RNA Polymerase II Transcription Pre-Initiation And Promoter Opening	10.0	2.13E-04	0.006	10.3	0.004	0.107
RNA Polymerase II Promoter Escape	10.0	2.13E-04	0.006	10.3	0.004	0.106
RNA Polymerase II HIV Promoter Escape	10.0	2.13E-04	0.006	10.3	0.004	0.103
↳Transcription of the HIV genome	7.7	1.92E-04	0.005	8.8	0.001	0.058
HIV Transcription Initiation	10.0	2.13E-04	0.006	10.3	0.004	0.101

Pathway (continued)	Aggressive SNPs Results			Non-aggressive SNPs Results		
	Fold Enrichment	Raw P value	FDR	Fold Enrichment	Raw P value	FDR
RNA Polymerase I Transcription Initiation	9.8	2.34E-04	0.006	3.4	0.261	1.740
↳RNA Polymerase I Promoter Clearance	6.4	0.001	0.026	2.2	0.369	2.190
↳RNA Polymerase I Transcription	7.5	2.20E-04	0.006	2.1	0.377	2.220
↳RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription	7.5	5.95E-06	4.23E-04	7.1	8.33E-04	0.040
TP53 Regulates Transcription of Cell Cycle Genes	9.6	2.56E-04	0.006	6.6	0.039	0.608
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression	9.5	0.001	0.022	< 0.01	1.000	3.570
↳Positive epigenetic regulation of rRNA expression	8.1	1.43E-04	0.004	4.7	0.072	0.850
↳Epigenetic regulation of gene expression	7.6	5.56E-06	4.10E-04	5.8	0.006	0.145
PPARA activates gene expression	9.2	7.66E-08	1.17E-05	8.6	9.24E-05	0.010
↳Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	9.9	9.54E-09	1.73E-06	8.5	1.02E-04	0.010
↳Fatty acid, triacylglycerol, and ketone body metabolism	4.9	1.02E-05	6.17E-04	4.2	0.003	0.105
RMTs methylate histone arginines	9.0	0.001	0.025	< 0.01	1.000	2.730
Factors involved in megakaryocyte development and platelet production	8.9	5.18E-10	1.47E-07	4.1	0.018	0.375
↳Hemostasis	2.4	0.002	0.038	1.1	0.789	3.650
TP53 Regulates Transcription of DNA Repair Genes	8.5	1.14E-04	0.004	7.3	0.009	0.216
Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants	8.2	4.94E-04	0.011	5.7	0.052	0.707
↳Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer	8.2	4.94E-04	0.011	5.7	0.052	0.702
↳Signaling by NOTCH1 in Cancer	8.2	4.94E-04	0.011	5.7	0.052	0.693
Constitutive Signaling by NOTCH1 PEST Domain Mutants	8.2	4.94E-04	0.011	5.7	0.052	0.697
↳Signaling by NOTCH1 PEST Domain Mutants in Cancer	8.2	4.94E-04	0.010	5.7	0.052	0.688
NoRC negatively regulates rRNA expression	8.0	1.54E-04	0.005	2.3	0.357	2.150
↳Negative epigenetic regulation of rRNA expression	7.7	1.92E-04	0.006	2.2	0.369	2.200
RNA Polymerase II Pre-transcription Events	7.3	2.52E-04	0.006	8.3	0.002	0.064
Senescence-Associated Secretory Phenotype (SASP)	6.3	0.002	0.027	4.3	0.081	0.903
Unclassified	0.7	3.98E-07	5.29E-05	0.8	0.005	0.129
GPCR downstream signaling	0.1	3.90E-04	0.009	< 0.01	0.003	0.104
↳Signaling by GPCR	0.2	6.98E-04	0.014	0.12	0.005	0.134
Transmembrane transport of small molecules	< 0.01	0.002	0.027	< 0.01	0.038	0.590

Genes used in the gene list analysis based on the PRSA SNPs included the following:

AIRE, AP-2, AP-2alpha, AP-2gamma, AR, ARID3A, ATF2, ATF3, Arnt, BATF, BCL11A, BCL3, BCL6, BCLAF1, BHLHE40, BLIMP1, Bcl6b, C/EBPbeta, C/EBPdelta, CBX3, CCNT2, CDX2, CEBPB, CEBPD, CHD1, CHD2, CKROX, COUP-TF=HNF-4, COUPdirectrepeat1, CREB1, CTCF, CUX1, Cdc5, Cdx, Cdx1, Cdx2, E2F, E2F-1, E2F-1:DP-1, E2F-1:DP-2, E2F-4:DP-2, E2F1, E2F3, E2F4, E2F6, EBF, EBF1, EGR1, ELF1, ELK1, ELK4, EP300, ESR1, ESRRA, ETS1, EWSR1-FLI1, Elf3, Evi-1, Evi1, FAC1, FOS, FOSL2, FOXA1, FOXA2, FOXB1, FOXC1, FOXC2, FOXD2, FOXD3, FOXJ2, FOXJ3, FOXM1, FOXO1, FOXO3A, FOXP1, FOXP2, FOXfactors, Foxa2, Foxd3, Foxj3, Foxk1, Foxl1, Freac-7, GABPB1, GAF, GATA-1, GATA-2, GATA-3, GATA-6, GATA1, GATA2, GATA3, GATA5, GATA6, GSX2, GTF2F1, GZF1, Gata1, Gata3, Gfi, Gfi-1, Gfi1, Gfi1b, HDAC1, HDAC2, HEY2,

HFH4(FOXJ1), HFH8(FOXF1A), HMGN3, HNF3, HNF4, HNF4A, HNF4alpha1, HOXA10, HOXA13, HOXC10, HOXC11, HSF1, Hnf4a, Hoxa10, Hoxa3, Hoxa9, Hoxb13, Hoxb6, Hoxb8, Hoxb9, Hoxc13, Hoxd10, IKZF1, IPF1, IRF-1, IRF-2, IRF1, ISGF-3, Ik-2, Irf3, Irx4, JUN, JUNB, JUND, KAISO, KDM5A, KDM5B, KROX, Klf7, LUN-1, MAFF, MAFK, MAX, MAZ, MAZR, MBD4, MEF-2, MEF2A, MEOX2, MTA3, MTF-1, MXI1, MYBL2, MYC, MZF1, Mafb, Mtf1, NF-1, NF-Y, NF-kappaB, NFATC1, NFE2L2, NFIC, NFKB1, NFYA, NFYB, NR1H2::RXRA, NR2C2, NR2F1, NR2F2, NR2F6, NR3C1, Nanog, Nkx2-5, ONECUT3, Oct-1, Olf-1, P50:P50, PHF8, PML, POLR2A, POU2F2, PPAR, PPAR=HNF-4=COUP=RAR, PPARG::RXRA, PPARalpha:RXRalpha, PPARdirectrepeat1, PPARgamma:RXRalpha, Pitx2, Plagl1, Pou4f3, RAD21, RARA, RARB, RARG, RBBP5, RCOR1, REST, RFX3, RFX5, RREB-1, RREB1, RUNX3, RXR:LXR-beta, RXRA, RXRB, RXRG, Rfx3, Rxra, SAP30, SCRT1, SCRT2, SETDB1, SIN3A, SMARCA4, SMARCB1, SMARCC1, SMARCC2, SMC3, SP1, SP1:SP3, SP4, SPI1, SPIB, SPIC, SREBP, SRF, STAT1, STAT3, STAT3:STAT3, STAT6, Six6, Smad3, Sox1, Sox11, Sox12, Sox15, Sox2, Sox4, Sox7, Sox8, Sp1, Sp3, Sp4, Srf, TAF1, TAF7, TAL1, TBP, TCF11:MafG, TCF12, TCF7, TCF7L2, TEAD1, TEAD3, TEAD4, TEF, TFAP2A, TFAP2C, TRIM28, Tcf1, Tcf2, Tcfap2a, Tcfap2c, Tcfap2e, Tcfe2a, UBTF, UF1H3BETA, USF1, USF2, WRNIP1, WT1, YY1, YY2, ZBRK1, ZBTB33, ZBTB7A, ZEB1, ZIC1, ZIC3, ZIC4, ZNF143, ZNF217, ZNF219, ZNF263, ZNF515, ZNF740, Zbtb3, Zec, Zfp105, Zfp281, Zfp423, Zfp740, Zfx, Zic1, Zic2, Zic3, Zscan4, aMEF-2, c-Myc:Max, mTERF

Genes used in the gene list analysis based on the PRS_{NA} SNPs included the following: (SATB1)2, AIRE, AP-1, AP-4, AR, ARID3A, ASCL2, ATF2, Ascl2, BCL6, BHLHE40, BLIMP1, Bach1, Bcl6b, C/EBP, C/EBPbeta, CAC-bindingprotein, CCNT2, CDPCR3, CEBPA,

CEBPB, CHD1, COUPTF, CREB, CREB1, CTCF, Cdx, Crx, DEAF1, DMRT1, DMRT2, DMRT3, DMRT5, Dbx1, Dbx2, Dlx4, Dmbx1, E2F6, EBF1, ELF5, EN1, EP300, ESRRA, EWSR1-FLI1, EZH2, Elf3, Elf5, Evi-1, FOXA1, FOXA2, FOXC1, FOXC2, FOXD3, FOXG1, FOXJ2, FOXJ3, FOXM1, FOXO3A, FOXP1, FRA1, Foxa2, Foxd3, Foxj3, Foxk1, Foxl1, Freac-7, GATA1, GATA2, GBX2, Glis2, HFH3(FOXI1), HMGIY, HNF1, HNF1A, HNF3, HNF4A, HNF4directrepeat1, HSF1, Hbp1, HeliosA, Hic1, Hoxb13, Hoxb3, Hoxd10, Hoxd13, IKZF1, INSM1, IRF4, LRF, LUN-1, MAFK, MAX, MEF-2, MEF2A, MEF2C, MEOX2, MXI1, MYC, MYF, Mafk, Mtf1, Myf, Myf6, NF-Y, NFATC1, NFIC, NFKB1, NR2F2, NeuroD, Nkx2-3, Nkx5-2, ONECUT1, ONECUT2, ONECUT3, Obox1, Obox5, Oct-1, Oct-6, Otx2, PLAG1, PML, POLR2A, POU1F1, POU2F1, POU3F3, POU4F1, POU4F2, POU4F3, POU5F1, PU.1, Pax-5, Pax-6, Pax6, Pbx1, Pou1f1, Pou3f1, Pou3f3, Pou3f4, Pou4f3, RAD21, RARA, RARB, RARG, REST, RORalpha, RREB1, RSRFC4, RUNX3, SMC3, SPDEF, SPI-B, SPI1, SPIC, SREBP, SREBP1, SRF, STAT5A, SUZ12, Sfpi1, Six6, Sox1, Sox11, Sox4, Spic, Srf, Staf, TAF1, TATA, TBP, TCF12, TCF7L2, TEAD1, TEAD3, TEAD4, TFAP2B, TP73, TRIM28, Tbp, Tcfap2e, ZNF143, Zfp105, Zfp128, Zfp161, mTERF

Table 1.6. Sensitivity analysis evaluating impact of FDR significance versus Bonferroni corrected significance in SNP selection procedure.

Note: The $PRS_{FDR,A}$ included the following ten SNPs: rs7767188, rs7968403, rs8102476, rs61890184, rs6763931, rs684232, rs7153648, rs12665339, rs1283104, and rs4924487. The $PRS_{FDR,NA}$ included the following 30 SNPs: rs2735839, rs9364554, rs878987, rs445114, rs7611694, rs17021918, rs34579442, rs10845938, rs1218582, rs1983891, rs28441558, rs11649743, rs17694493, rs76934034, rs7127900, rs12500426, rs6465657, rs13252298, rs7000448, rs12785905, rs5945572, rs12543663, rs9623117, rs2807031, rs11650494, rs2660753, rs33984059, rs2273669, rs59308963, and rs1270884.There was no overlap in the selected aggressive SNPs between FDR and Bonferroni correction approaches but there was overlap in the selected non-aggressive SNPs.

Case-only Analysis	**$PRS_{FDR,A}$**				**$PRS_{FDR,NA}$**			
	Ratio (95% CI)			**P-value**	**Ratio (95% CI)**			**P-value**
Non-Hispanic Whites								
Composite Aggressive	1.11	(1.05,	1.18)	1.71E-04	0.87	(0.82,	0.92)	7.60E-07
Early Onset	1.23	(1.09,	1.39)	6.29E-04	1.10	(0.97,	1.24)	0.137
High Gleason	1.03	(0.94,	1.12)	0.523	0.84	(0.77,	0.92)	1.46E-04
Biochemical Recurrence	1.17	(1.10,	1.26)	2.77E-06	0.89	(0.83,	0.95)	4.94E-04
African Americans								
Composite Aggressive	1.20	(1.00,	1.45)	0.052	0.95	(0.79,	1.14)	0.581
Early Onset	1.17	(0.91,	1.51)	0.234	1.30	(1.02,	1.67)	0.039
High Gleason	1.34	(1.02,	1.77)	0.037	0.94	(0.72,	1.23)	0.637
Biochemical Recurrence	1.17	(0.94,	1.44)	0.155	1.06	(0.86,	1.30)	0.605

Table 1.7. Sensitivity analysis evaluating impact of insufficient follow-up time for determining BCR.

Case-only Analyses	PRS_{167}			PRS_A			PRS_{NA}		
	Ratio (95% CI)		P-value	Ratio (95% CI)		P-value	Ratio (95% CI)		P-value
Non-Hispanic Whites									
Composite Aggressive	0.99	(0.93, 1.05)	0.662	1.15	(1.09, 1.22)	1.46E-06	0.90	(0.85, 0.95)	3.65E-04
Biochemical Recurrence	1.02	(0.95, 1.09)	0.655	1.13	(1.05, 1.21)	4.82E-04	0.93	(0.87, 1.00)	0.039
African Americans									
Composite Aggressive	1.00	(0.83, 1.20)	0.983	1.27	(1.06, 1.52)	0.010	1.03	(0.85, 1.24)	0.795
Biochemical Recurrence	1.03	(0.83, 1.27)	0.780	1.19	(0.97, 1.45)	0.095	0.98	(0.79, 1.22)	0.877

Figure 1.1. Manhattan plots from the genome-wide aggressive case-control analysis.

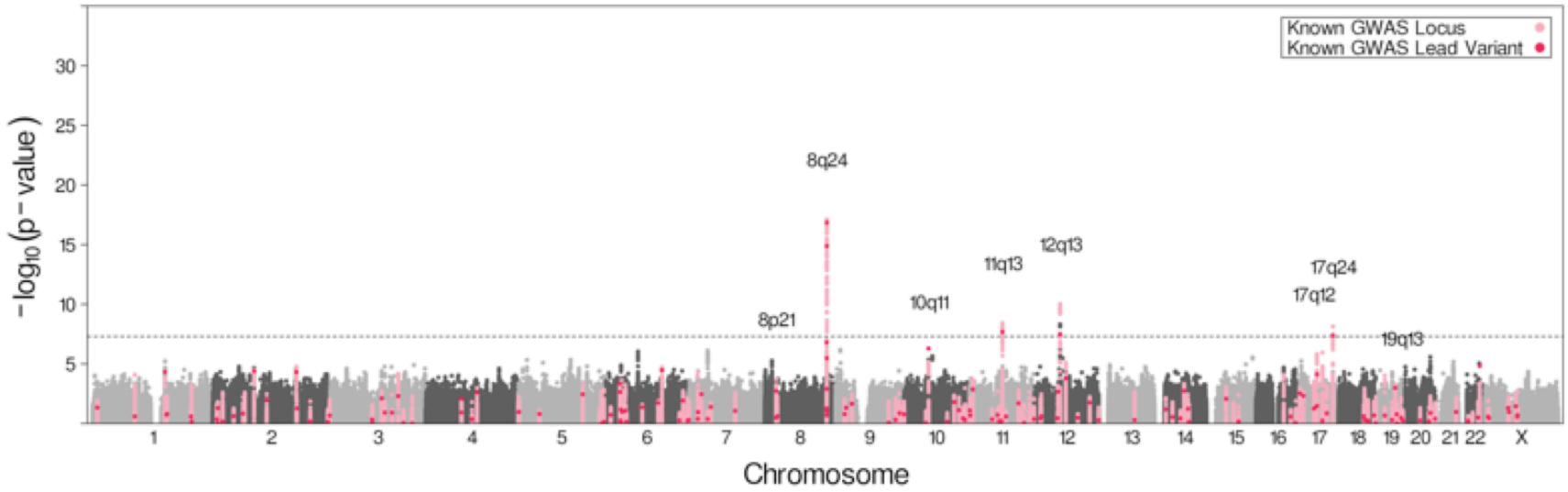


Figure 1.2. Manhattan plots from the genome-wide non-aggressive case-control analysis.

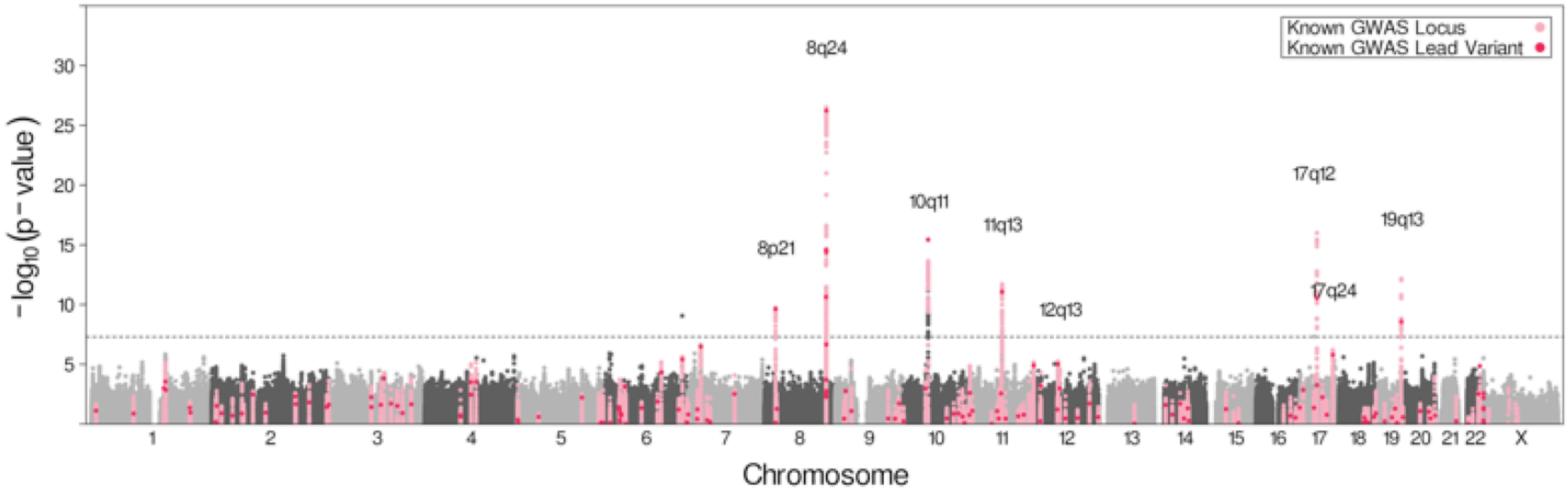


Figure 1.3. Manhattan plots from the genome-wide case-only analysis.

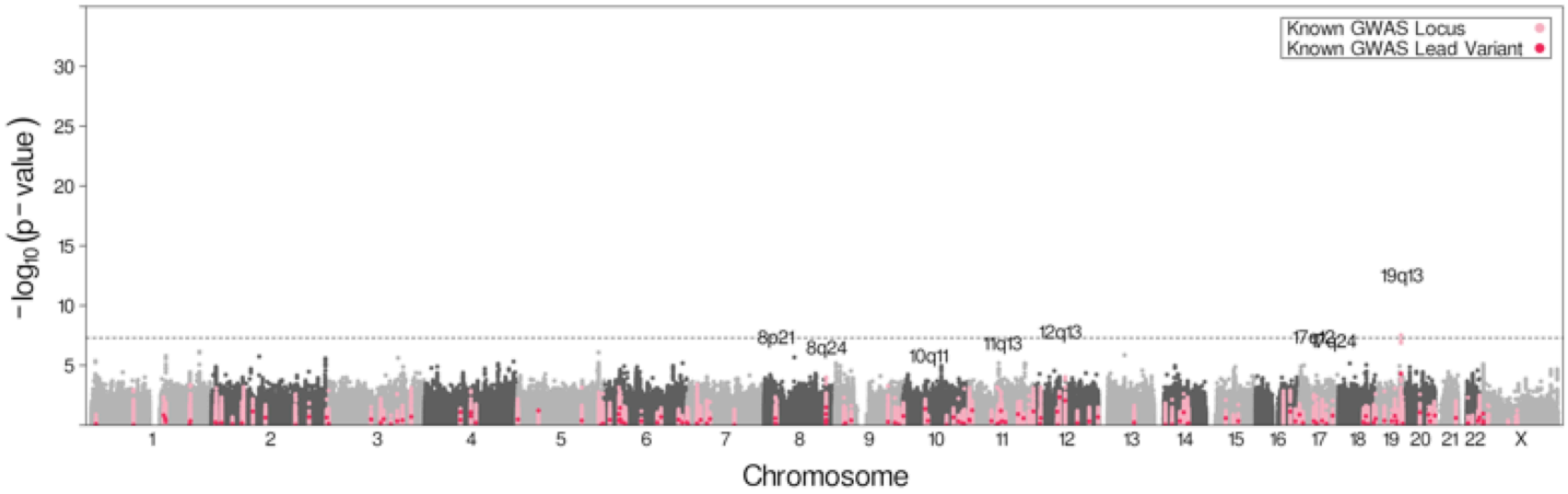
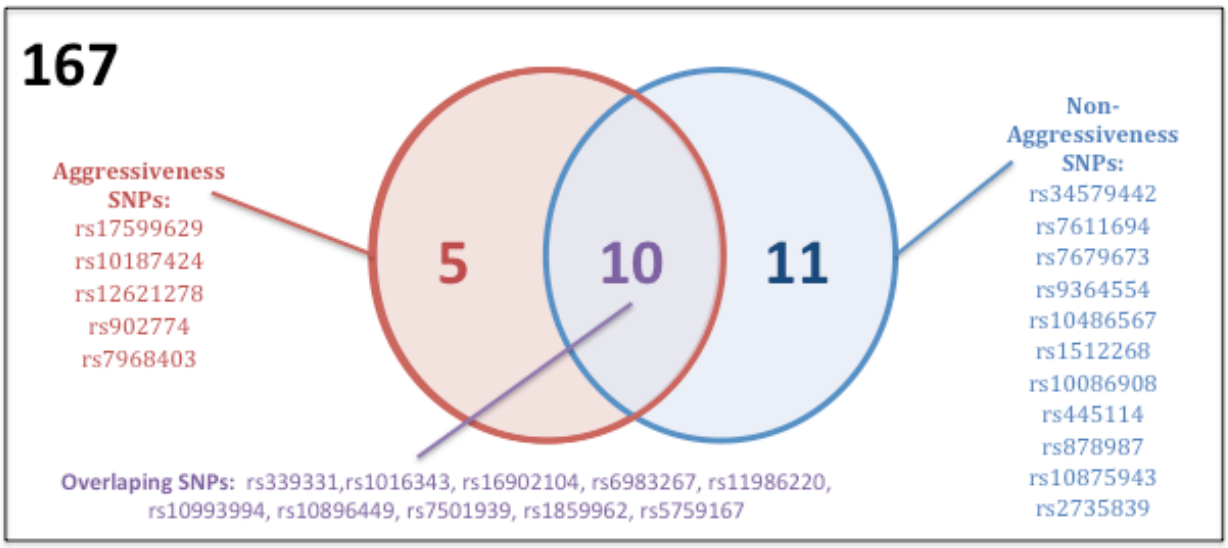


Figure 1.4. Aggressive versus non-aggressive SNPs among the 167 known prostate cancer loci

Legend: The Venn diagram in Figure 4 depicts the three subsets of SNPs among the 167 known Prostate Cancer loci that are either associated only with aggressive disease (red), associated only with non-aggressive disease (blue) or associated with both (purple).



Supplemental Table 1.1. List of the 167 known prostate cancer loci

SNP	Region	Chr	Position	Risk Allele	Reference Allele	Previously published Odds Ratio	Source
rs636291	1p35	1	10556097	A	G	1.18	Han 2015
rs56391074	1p22	1	88210715	A	AT	1.05	Schumacher 2018
rs17599629	1q21	1	150658287	G	A	1.08	Han 2015
rs34579442	1q21	1	153899900	C	CT	1.07	Schumacher 2018
rs1218582	1q21	1	154834183	G	A	1.06	Han 2015
rs4245739	1q32	1	204518842	A	C	1.10	Han 2015
rs1775148	1q32	1	205757824	C	T	1.06	Han 2015
rs62106670	2p25	2	8597123	T	C	1.05	Schumacher 2018
rs11902236	2p25	2	10117868	A	G	1.07	Han 2015
rs9287719	2p25	2	10710730	C	T	1.06	Han 2015
rs13385191	2p24	2	20888265	G	A	1.07	Han 2015
rs1465618	2p21	2	43553949	A	G	1.08	Han 2015
rs721048	2p15	2	63131731	A	G	1.15	Han 2015
rs74702681	2p14	2	66652885	T	C	1.17	Schumacher 2018
rs10187424	2p11	2	85794297	A	G	1.09	Han 2015
rs11691517	2q13	2	111893096	T	G	1.07	Schumacher 2018
rs12621278	2q31	2	173311553	A	G	1.33	Han 2015
rs34925593	2q31	2	174234547	C	T	1.05	Schumacher 2018
rs59308963	2q33	2	202123479	T	TATTCTGTC	1.05	Schumacher 2018
rs2292884	2q37	2	238443226	G	A	1.14	Han 2015
rs3771570	2q37	2	242382864	A	G	1.12	Han 2015
rs2660753	3p12	3	87110674	T	C	1.13	Han 2015
rs2055109	3p11	3	87467332	C	T	1.20	Han 2015
rs1283104	3q13	3	106962521	G	C	1.05	Schumacher 2018
rs7611694	3q13	3	113275624	A	C	1.10	Han 2015
rs10934853	3q21	3	128038373	A	C	1.12	Han 2015
rs6763931	3q23	3	141102833	T	C	1.04	Han 2015
rs182314334	3q25	3	152004202	T	C	1.09	Schumacher 2018
rs142436749	3q26	3	169093100	G	A	1.25	Schumacher 2018
rs10936632	3q26	3	170130102	A	C	1.11	Han 2015
rs10009409	4q13	4	73855253	T	C	1.08	Han 2015
rs1894292	4q13	4	74349158	G	A	1.10	Han 2015
rs12500426	4q22	4	95514609	A	C	1.08	Han 2015
rs17021918	4q22	4	95562877	C	T	1.11	Han 2015
rs7679673	4q24	4	106061534	C	A	1.10	Han 2015
rs2242652	5p15	5	1280028	G	A	1.15	Han 2015
rs12653946	5p15	5	1895829	T	C	1.10	Han 2015
rs2121875	5p12	5	44365545	G	T	1.05	Han 2015
rs10793821	5q31	5	133836209	T	C	1.05	Schumacher 2018
rs76551843	5q35	5	169172133	A	G	1.31	Schumacher 2018
rs6869841	5q35	5	172939426	A	G	1.07	Han 2015
rs4976790	5q35	5	177968915	T	G	1.08	Schumacher 2018
rs4713266	6p24	6	11219030	C	T	1.06	Han 2015
rs115457135	6p22	6	30073776	A	G	1.07	Han 2015
rs12665339	6p21	6	30601232	G	A	1.06	Schumacher 2018
rs130067	6p21	6	31118511	G	T	1.05	Han 2015
rs3096702	6p21	6	32192331	A	G	1.07	Han 2015
rs115306967	6p21	6	32400939	G	C	1.06	Han 2015
rs9296068	6p21	6	32988695	T	G	1.05	Schumacher 2018
rs9469899	6p21	6	34793124	A	G	1.05	Schumacher 2018
rs1983891	6p21	6	41536427	T	C	1.09	Han 2015
rs370814360	6p21	6	43694598	T	C	1.05	Schumacher 2018
rs9443189	6q14	6	76495882	G	A	1.08	Han 2015
rs2273669	6q21	6	109285189	G	A	1.07	Han 2015
rs339331	6q22	6	117210052	T	C	1.08	Han 2015
rs1933488	6q25	6	153441079	A	G	1.12	Han 2015
rs9364554	6q25	6	160833664	T	C	1.08	Han 2015
rs527510716	7p22	7	1944537	C	G	1.06	Schumacher 2018
rs11452686	7p21	7	20414110	T	TA	1.05	Schumacher 2018
rs12155172	7p15	7	20994491	A	G	1.11	Han 2015
rs10486567	7p15	7	27976563	G	A	1.19	Han 2015
rs17621345	7p14	7	40875192	A	C	1.07	Schumacher 2018
rs56232506	7p12	7	47437244	A	G	1.06	Han 2015
rs6465657	7q21	7	97816327	C	T	1.11	Han 2015

SNP	Region	Chr	Position	Risk Allele	Reference Allele	Previously published Odds Ratio	Source
rs2928679	8p21	8	23438975	T	C	1.05	Han 2015
rs1512268	8p21	8	23526463	A	G	1.18	Han 2015
rs11135910	8p21	8	25892142	A	G	1.11	Han 2015
rs12543663	8q24	8	127924659	C	A	1.08	Han 2015
rs10086908	8q24	8	128011937	T	C	1.15	Han 2015
rs1016343	8q24	8	128093297	T	C	1.25	Han 2015
rs13252298	8q24	8	128095156	A	G	1.19	Han 2015
rs6983561	8q24	8	128106880	C	A	1.47	Han 2015
rs116041037	8q24	8	128131809	A	G	2.45	Han 2015
rs445114	8q24	8	128323181	T	C	1.14	Han 2015
rs16902104	8q24	8	128340908	T	C	1.21	Han 2015
rs6983267	8q24	8	128413305	G	T	1.23	Han 2015
rs7000448	8q24	8	128441170	T	C	1.14	Han 2015
rs11986220	8q24	8	128531689	A	T	1.36	Han 2015
rs1048169	9p22	9	19055965	C	T	1.06	Schumacher 2018
rs17694493	9p21	9	22041998	G	C	1.08	Han 2015
rs10122495	9p13	9	34049779	T	A	1.05	Schumacher 2018
rs817826	9q31	9	110156300	C	T	1.41	Han 2015
rs1571801	9q33	9	124427373	A	C	1.07	Han 2015
rs1182	9q34	9	132576060	A	C	1.06	Schumacher 2018
rs141536087	10p15	10	854691	GCGCA	G	1.08	Schumacher 2018
rs76934034	10q11	10	46082985	T	C	1.13	Han 2015
rs10993994	10q11	10	51549496	T	C	1.23	Han 2015
rs1935581	10q23	10	90195149	C	T	1.05	Schumacher 2018
rs3850699	10q24	10	104414221	A	G	1.10	Han 2015
rs7094871	10q25	10	114712154	G	C	1.04	Schumacher 2018
rs2252004	10q26	10	122844709	G	T	1.16	Han 2015
rs4962416	10q26	10	126696872	C	T	1.09	Han 2015
rs1881502	11p15	11	1507512	T	C	1.06	Schumacher 2018
rs7127900	11p15	11	2233574	A	G	1.22	Han 2015
rs61890184	11p15	11	7547587	A	G	1.07	Schumacher 2018
rs547171081	11p11	11	47421962	CGG	C	1.05	Schumacher 2018
rs1938781	11q12	11	58915110	C	T	1.16	Han 2015
rs2277283	11q12	11	61908440	C	T	1.06	Schumacher 2018
rs12785905	11q13	11	66951965	C	G	1.12	Schumacher 2018
rs10896449	11q13	11	68994667	G	A	1.19	Han 2015
rs11290954	11q13	11	76260543	AC	A	1.06	Schumacher 2018
rs11568818	11q22	11	102401661	A	G	1.10	Han 2015
rs1800057	11q22	11	108143456	G	C	1.16	Schumacher 2018
rs11214775	11q23	11	113807181	G	A	1.07	Han 2015
rs138466039	11q24	11	125054793	T	C	1.32	Schumacher 2018
rs878987	11q25	11	134266372	G	A	1.07	Schumacher 2018
rs2066827	12p13	12	12871099	T	G	1.06	Schumacher 2018
rs10845938	12p13	12	14416918	G	A	1.06	Schumacher 2018
rs80130819	12q13	12	48419618	A	C	1.14	Han 2015
rs10875943	12q13	12	49676010	C	T	1.07	Han 2015
rs902774	12q13	12	53273904	A	G	1.17	Han 2015
rs7968403	12q14	12	65012824	T	C	1.06	Schumacher 2018
rs5799921	12q21	12	90160530	GA	G	1.06	Schumacher 2018
rs1270884	12q24	12	114685571	A	G	1.07	Han 2015
rs7295014	12q24	12	133067989	G	A	1.05	Schumacher 2018
rs9600079	13q22	13	73728139	T	G	1.01	Han 2015
rs1004030	14q11	14	23305649	T	C	1.05	Schumacher 2018
rs11629412	14q13	14	37138294	C	G	1.06	Schumacher 2018
rs8008270	14q22	14	53372330	G	A	1.12	Han 2015
rs7153648	14q23	14	61122526	C	G	1.11	Han 2015
rs7141529	14q24	14	69126744	G	A	1.09	Han 2015
rs8014671	14q24	14	71092256	G	A	1.06	Han 2015
rs4924487	15q15	15	40922915	C	G	1.06	Schumacher 2018
rs33984059	15q21	15	56385868	A	G	1.19	Schumacher 2018
rs112293876	15q22	15	66764641	C	CA	1.06	Schumacher 2018
rs11863709	16q21	16	57654576	C	T	1.16	Schumacher 2018
rs12051443	16q22	16	71691329	A	G	1.06	Han 2015
rs201158093	16q23	16	82178893	TAA	TA	1.05	Schumacher 2018
rs684232	17p13	17	618965	G	A	1.10	Han 2015
rs28441558	17p13	17	7803118	C	T	1.16	Schumacher 2018
rs142444269	17q11	17	30098749	C	T	1.07	Schumacher 2018

SNP	Region	Chr	Position	Risk Allele	Reference Allele	Previously published Odds Ratio	Source
rs11649743	17q12	17	36074979	G	A	1.15	Han 2015
rs7501939	17q12	17	36101156	C	T	1.22	Han 2015
rs11650494	17q21	17	47345186	A	G	1.15	Han 2015
rs7210100	17q21	17	47436749	A	G	1.51	Han 2015
rs2680708	17q22	17	56456120	G	A	1.05	Schumacher 2018
rs1859962	17q24	17	69108753	G	T	1.19	Han 2015
rs8093601	18q21	18	51772473	C	G	1.05	Schumacher 2018
rs28607662	18q21	18	53230859	C	T	1.08	Schumacher 2018
rs12956892	18q21	18	56746315	T	G	1.05	Schumacher 2018
rs533722308	18q21	18	60961193	CT	C	1.05	Schumacher 2018
rs10460109	18q22	18	73036165	T	C	1.05	Schumacher 2018
rs7241993	18q23	18	76773973	G	A	1.09	Han 2015
rs11666569	19p13	19	17214073	C	T	1.05	Schumacher 2018
rs118005503	19q12	19	32167803	G	C	1.09	Schumacher 2018
rs8102476	19q13	19	38735613	C	T	1.12	Han 2015
rs11672691	19q13	19	41985587	G	A	1.08	Han 2015
rs61088131	19q13	19	42700947	T	C	1.06	Schumacher 2018
rs2735839	19q13	19	51364623	G	A	1.15	Han 2015
rs103294	19q13	19	54797848	C	T	1.28	Han 2015
rs11480453	20q11	20	31347512	C	CA	1.05	Schumacher 2018
rs12480328	20q13	20	49527922	T	C	1.13	Han 2015
rs6091758	20q13	20	52455205	G	A	1.07	Schumacher 2018
rs2427345	20q13	20	61015611	G	A	1.06	Han 2015
rs6062509	20q13	20	62362563	A	C	1.12	Han 2015
rs1041449	21q22	21	42901421	G	A	1.06	Han 2015
rs2238776	22q11	22	19757892	G	A	1.08	Han 2015
rs9625483	22q12	22	28888939	A	G	1.14	Schumacher 2018
rs9623117	22q13	22	40452119	C	T	1.18	Han 2015
rs5759167	22q13	22	43500212	G	T	1.16	Han 2015
rs2405942	23p22	23	9814135	A	G	1.14	Han 2015
rs17321482	23p22	23	11482634	C	T	1.07	Schumacher 2018
rs5945572	23p11	23	51229683	A	G	1.23	Han 2015
rs2807031	23p11	23	52896949	C	T	1.07	Han 2015
rs5919432	23q12	23	67021550	A	G	1.06	Han 2015
rs6625711	23q13	23	70139850	A	T	1.04	Han 2015
rs4844289	23q13	23	70407983	G	A	1.04	Han 2015

Chapter 2

Detecting Gene-Environment Interactions in Human Birth Defects: Study Designs and Statistical Methods

Introduction

Abnormal embryonic development likely depends on both inherited genetic risk factors that reflect interactions between maternal and paternal genetics and on maternal environmental influences at specific gestational times. Such gene-environment interactions (GxE) have been reported for several congenital abnormalities, including neural tube defects (NTDs) (1), congenital heart defects (CHD) (2), and oral–facial clefts (3). Nevertheless, the biological mechanisms underlying such interactions generally remain somewhat unclear, and public health recommendations have only begun to consider how combinations of genetic and environmental effects could interact to influence the risk of birth defects (4).

Often GxE studies build upon existing knowledge about the individual main effects of genetic or environmental factors. For example, animal studies, observational research, and clinical trials determined the existence of a link between folate and the occurrence of NTDs (5)(6–9)(10,11). Later, Christensen et al. (12) proposed that there was also a gene-nutrient interaction between a folate related gene, methylenetetrahydrofolate reductase (*MTHFR*), and maternal folate levels for NTDs. Furthermore, we are often interested in scenarios where the GxE effect is greater than the simple combination of the genetic or environmental factor alone. Before widespread availability of genotyped genetic data, family history was used as a proxy measurement for inherited genetic risk. For example, (13) conducted a case-control study for clubfoot that found the joint effect of smoking and family history (odds ratio (OR) = 20.30, 95%

confidence interval (95%CI): 7.90 to 52.17) to be greater than the individual or multiplicative effects of either smoking (OR = 1.34, 95%CI: 1.04, 1.72) or family history (OR = 6.52, 95% CI: 2.95, 14.41).

In GxE studies of birth defects, it is important to recognize that multiple interaction effects may occur simultaneously. Any maternal environmental exposure may interact with either the mother's genes or the infant's genes to alter risks for birth defects, resulting in coexisting infant GxE and maternal GxE effects. Additionally, there could be interactions between infants' genotype and their mothers' genotype, yielding maternal-fetal gene-gene interactions (GxG).

Researchers can investigate such interactions in the National Birth Defects Prevention Study (NBDPS), a population based case-control study with 10 recruitment centers across the country (Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas, and Utah). The NBDPS collected DNA samples from over 19,000 stillborn and live-born case infants with birth defects from a pre-specified list of over 30 conditions. Also available is information on the mothers of these case infants regarding genetic, lifestyle and environmental exposures, as well as genetic data on fathers who provided DNA samples (14,15). The same information was collected from over 6,000 population-based unmatched control infants without birth defects and their parents (16).

The NBDPS has detected a number of associations, including between birth defects and maternal "environmental" factors such as obesity (17) and diet quality (18) or with mothers' pharmaceutical drug usage of antidepressants (19) and nitrosatable drugs (20). Investigators have also found associations between candidate genes and birth defects including genes in folate-related pathways for CHD (21) and glucose homeostasis genes for NTDs (22). (2) also identified

multiple gene-environment interactions between folate-related genes and maternal smoking or alcohol intake for CHDs. These examples represent only a handful of the numerous etiologic questions that either have or could be investigated in the NBDPS.

There are many considerations for epidemiologic studies in which the main research objective is to interrogate GxE associations for birth defects. These include choices in study design, type of genetic information to collect, and statistical approaches to use. Here we present and discuss some of the current study designs for, analytic methods used in, and challenges to identifying GxE effects in studies of birth defects.

Methods

Measuring Genes and Environment

Genes. To date, most studies of GxE for birth defects have evaluated single-nucleotide polymorphisms (SNPs) from candidate genes or loci. SNPs are generally chosen based on some prior biological rationale or from main effect findings from candidate gene studies. For example, in gene-nutrient studies, only genes from a pathway involved in the transport or metabolism of that nutrient may be considered.

The focus on candidate genes has given way to genome-wide assessments of GxE, on account of the increasing affordability of high-resolution genotyping arrays that are commonly used in genome wide association studies (GWAS). Genome-wide interrogations of GxE effects, termed gene-by-environment wide interaction studies (GEWIS) (23) often begin with individually testing each SNP for its association with the outcome of interest, and then the most strongly associated SNPs are evaluated for their interactions with environmental exposures. (24) developed a method to extend GWAS analyses to include a GEWIS analysis that uses a two-step approach. The first step serves as a screening phase and employs a test similar to the case-only

approach (described below). In the second step, only SNPs that are significant in the first step are tested for GxE using the traditional case-control approach.

In many two-step approaches investigators must optimize filtering significance (i.e., p-values) thresholds in their selection of SNPs for analysis in the second step. Alternatively, one can use two-stage data collection with high-density genotyping and analysis in the first stage followed by genotyping of only the most statistically significant SNPs using a custom array in the second stage (25,26). However the cost differential between high-density and targeted arrays has become small enough that most studies simply run a high-density array on all study subjects. Several Bayesian approaches are also available that average results from both case-only and case-control approaches to estimate GxE effects (26). Specific pathway based analyses combine the strengths of candidate gene and GWAS approaches; inter-related genes that play a part in the same biological pathway are considered while using genetic data from genome-wide genotyping arrays. This approach has only been recently applied to GxE studies (27).

One major point of contention in all GxE studies is whether detection of a main genetic effect is necessary prior to testing interaction effects. In birth defects research, GxE effects are typically assessed only for variants that have already shown evidence of a main genetic effect. Recent methods, however, suggest that this approach may overlook important findings. For example, some epidemiologic studies may have been unable to detect a clear association between the C677T allele in the *MTHFR* gene, which plays a role in folate metabolism, and risk of NTDs due to a lack of varying levels of folate intake in some study populations, especially post fortification. It was only after a meta-analysis was conducted using 17 studies from North America and Europe that the link was finally confirmed (28). Therefore, SNPs with a GxE effect may not make it to the list of candidate SNPs due to a diminished genetic effect in the absence of

an environmental exposure that also contributes to the outcome, as was the case with the *MTHFR* variant and folate intake for NTDs (Daly et al.,1995; Crider et al., 2014). Some SNPs may only be associated with an outcome in the presence of an environmental factor. In such a case, the main effect for a SNP will not be found for a causal reason—because both the genetic factor and the environmental factor must be present to observe the outcome. (1) reported that the risk of NTDs was only slightly elevated for infants who possessed a risk SNP, rs11627387, in another folate-related gene *MTHFD1* (OR=1.11, 95%CI: 0.87 to 1.41), but for infants who also had low folate intake the risk of NTDs increased four-fold (OR=4.25, 95%CI: 2.33 to 7.75). If there is a true etiologic mechanism by which the genetic and environmental factors interact to produce the outcome, then a main genetic effect should not be required to investigate the presence of a GxE effect.

Environment. There are unique considerations in studying maternal exposures just prior to and during early pregnancy. In the data collection phase, maternal environmental exposure ascertainment is challenging because mothers often underreport behaviors or lifestyle choices that are known to potentially cause harm to their fetuses. A study in New Zealand found that maternal smoking was underreported for half of all mothers based on a comparison between self-reported smoking and levels of serum cotinine, a nicotine metabolite (29). Additionally, mothers that did not respond to the lifestyle questionnaire were more likely to be heavy smokers in the first trimester (40%) than mothers who did respond (16%) (29). Similar underreporting has also been found for medication use in the United States (30). In addition, other environmental exposures may be difficult to accurately measure, such as pollutants or industrial byproducts (e.g., Bisphenol A), although for many of these, any biased recall may not be differ between cases and controls. Since many single birth defect phenotypes are rare, most research participants

are identified after the outcome has occurred and are asked to retrospectively recall their exposures during narrow windows of time. At times women may be asked about exposures that occurred many months or years earlier.

For main environmental effects, if the level of misclassification is similar (non-differential) among cases and controls, then estimates for the environmental factor will be biased towards the null, making it more difficult to detect any true effects. If the misclassification is more (or less) prevalent among cases than controls (differential), then estimates can be biased in either direction. These issues can become more complex in the context of GxE estimates and are less often discussed (31). The direction of for a GxE effect estimate may be predicted when there is no true association between the genotype and environmental exposure among the controls, and when misclassification of the environmental exposure is non-differential between genotypes (García-Closas et al., 1998). Under these conditions, in the presence of a true multiplicative interaction, differential misclassification of the environmental exposure biases the GxE effect estimate toward the null. In the absence of multiplicative interaction, differential misclassification of the environmental exposure does not bias the GxE effect estimate. Importantly, the first of these conditions could be tested in the NBDPS data via a test of independence of the environmental and genotype data in the controls.

Design and Statistical Analysis for Gene-Environment Interactions

The following discussions of study designs and analyses will be limited to dichotomous phenotypes since most birth defect phenotypes are defined as binary. Similarly we will assume a qualitative (binary) environmental exposure but many approaches can be extended to quantitative (continuous) environmental exposures.

Case-control studies of unrelated individuals are commonly used for evaluating rare diseases such as birth defects. However, case-only (i.e., affected infants), case-parent dyad (i.e., affected infants and their mothers), and triad (i.e., affected infants and both of their parents) designs can also be used to evaluate GxE. Note that the case-sibling study design in which an affected individual is matched to his or her sibling can be more efficient for detecting GxE than case-control or triad designs especially for rare SNPs (Witte et al., 1999; Chatterjee et al., 2005). We will not expand on the discordant sibship model because our goal here is to describe the possible approaches that could be applied to data from the NBDPS, which contains affected and unaffected family dyads and triads, and does not include siblings.

In general, sample sizes required to detect GxE can be substantially larger than those required for detecting main genetic or environmental effects. They are influenced by: 1) the magnitude of the GxE effect; 2) the allele frequency of the causal SNP; 3) the prevalence of the environmental exposure in the study population; 4) the model of inheritance assumed; and 5) the strength of correlation between causal and tag SNPs. Strong main genetic effects could also affect the ability to detect GxE, although the extent of this depends on the causal structure of the GxE relationship (Thomas, 2010). To get a sense of the sample sizes required, assuming a true GxE odds ratio of 1.25, an environmental exposures with 40% prevalence, a minor allele frequency of 0.3, a dominant mode of inheritance and directly measuring the causal SNPs requires over 10,000 case-control pairs to achieve 80% power (Hein et al., 2008). In contrast, if the GxE odds ratio is 2.0, fewer than 1,600 pairs would be required, holding everything else constant. In many power and sample size calculations, a larger GxE effect is typically expected. For example, an assumption of at least a 3-fold increased risk under a gene-environment interaction model is consistent with, and is in fact less than, what others have used as a guide

(Hwang et al., 1994). Although Hein et al. (2008) found that sample size requirements did not differ greatly based on the assumed inheritance model (e.g., recessive, additive, or dominant), typically recessive models require greater sample sizes to achieve the same level of power than dominant models in case-only (Clarke and Morris, 2010) and case-control (Palmer and Cardon, 2005) studies. Since the causal SNP involved in the GxE effect is generally not directly genotyped but rather assayed by a tag SNP, the degree of linkage disequilibrium and concordance of allele frequencies between the causal and tag SNPs among the controls also impact sample size requirements. There is an inverse relationship between the causal and tag SNP correlation and the sample size required to detect GxE effects would increase (Hein et al., 2008).

There is specialized statistical software available for GxE power calculations, including Quanto (Gauderman, 2002) for matched case-control, case-sibling, case-parent, and case-only study designs, PBAT (32) or the R implementation, pbatR, (33) for family-based study designs, and the R package, trio, (Schwender et al., 2014) specifically for trios. A set-based approach is implemented in SBERIA, which weights main effects of SNPs within a gene region using a genetic risk score that is included in interaction tests (Jiao et al., 2013).

Results

Case-Control Study. The case-control approach utilizes information from affected and unaffected individuals who are unrelated as depicted in **Figure 1**. To evaluate GxE in a case-control study, the traditional approach is to fit a logistic regression model with terms that include the genetic effect, environmental effect, GxE effect, and any covariates:

$$\text{Logit}(P[Y=1 \mid G, E, C]) = \beta_0 + \beta_G*(G) + \beta_E*(E) + \beta_{GxE}*(G*E) + \beta_C *C \quad (1)$$

The vector Y represents the binary outcome with possible values for affected (1) and unaffected individuals (0). The model intercept is given by β_0, while G represents the genotype of the affected or unaffected individuals following an additive, dominant or recessive model and its effects on Y are estimated by β_G. The parameter E is the maternal environmental exposure of the affected or unaffected individual and its effect is estimated by β_E. The GxE effect is estimated by β_{GxE}. Therefore evaluating the presence of an interaction between the genetic and environmental factors requires testing the null hypothesis that $\beta_{GxE} = 0$ (i.e., no interaction), while $\beta_{GxE} \neq 0$ indicates interaction is present. A joint two-degrees of freedom test of both the main genetic (β_G) and GxE (β_{GxE}) effects can provide greater power than standard case-control and case-only approaches when the causal structure describing genetic, environmental, and GxE effects on the outcome is not well known (Kraft, 2007). The case-control approach is common in genetic association studies. As an example of a basic GxE test, in a population-based case-control study of 69 infants with cleft palate and 284 controls with non-cleft birth defects, Hwang et al. (1994) observed a GxE effect between maternal smoking and the infant's genotype for the transforming grown factor alpha (*TGFA*) locus.

The case-control design is susceptible to population stratification bias, especially when the study includes individuals with varying genetic ancestries, as is the case with the NBDPS. The issue arises if both the frequency of genetic variants and the risk of birth defects vary among the ancestrally different populations. This can result in a non-causal association between genetic ancestry and the individual's outcome even in the absence of a true association between genotypes and birth defect phenotypes due to confounding. This is generally adjusted for in the logistic regression model by including covariates (C) as a vector of eigenvectors, which represent the principal components that identify distinct subpopulations or by including ancestry

informative markers. Their effects are then estimated by β_C. Commonly used methods to generate these principal components include EIGENSTRAT using the EIGENSOFT package (Patterson et al., 2006; Price et al., 2006) and multi-dimensional scaling using PLINK (Purcell et al. 2007). The case-control logistic regression model can be fit in any standard statistical software package, such as in R using the glm function, with the CGEN package (34) or in PLINK (35) as described in **Table 1**.

Case-Only Study. This design requires only affected individuals (**Figure 1**). The case-only approach can be powerful for detecting GxE, and useful when identifying and recruiting control subjects is problematic or not feasible. However, the approach assumes that the genetic and environmental factors are independent in the study's source population (36). One can estimate the association between gene and environment among cases from the model using logistic regression:

$$\text{Logit}(P[G = 1 \mid E]) = \beta_0 + \beta_{GxE} * (E), \quad (2)$$

where G is the genotype (estimated for an dominant or recessive model in this case, but could be coded for an additive model using ordinal or multinomial logistic models (Clarke and Morris 2010)), E is the environmental exposure, β_0 is the model intercept and β_{GxE} is the ln(OR) for the association between environment and genotype (i.e., the estimate of the GxE effect). This OR among cases is equivalent to the synergy index calculated from a case-control design, which is the OR of the GxE effect divided by the product of the OR of the genetic effect and the OR of the environmental exposure (37). A synergy index greater than one implies the presence of multiplicative GxE. This approach could be especially useful in re-analyzing existing data using only the cases of prior case-control studies. As an example of this study design, Zeiger et al. (2005) implemented a case-only pooled meta-analysis using 335 cleft palate cases from five

previously published studies and found a GxE effect between the *TGFA* Taq1 C2 allele genotype and maternal smoking.

The case-only analysis can be more powerful (and potentially, financially much less costly if data collection of controls is required) than a traditional case-control study, as the assumption of gene-environment independence reduces the standard error of the estimate of GxE interaction (Khoury and Flanders, 1996; Albert, 2001). Violations of the independence assumption, however, would preclude using the case-only design to assess GxE (38). The independence assumption may not hold when behavior is affected by genetic factors (39). For example, those who possess the risk allele in the D2 dopamine receptor gene (*DRD2*) may experience increased dopamine levels when consuming nicotine and thus more likely to be cigarette smokers (40). A case-only study of the interaction between *DRD2* and smoking in lung cancer could be problematic if this relationship between *DRD2* and smoking existed in the population. Note that many of the family-based methods (discussed below) also require that gene and environment are independent within families (41,42).

Potential drawbacks to the case-only approach are that main effects (genetic or environmental) cannot be estimated, and it is vulnerable to bias due to population stratification (43) (**Table 1**). However, if truly causal, the presence of both genetic and environmental factors is more likely to occur in cases than controls, which makes this approach appealing for studying rare genes or environmental exposures. The case-only logistic regression analysis can be implemented using standard statistical software.

Family-Based Trio Study. GxE can also be evaluated with a study of parents and an affected infant (**Figure 1**). The most common approach here is the transmission-disequilibrium test (TDT). The affected infant is the case (with genotypes transmitted from the parents), and is

matched to three pseudo-controls (with the non-transmitted parental genotypes). One then conducts what is essentially a matched 1:3 case-control analysis where each family serves as a matching stratum (44,45). This approach requires genetic information from both affected individual's parents; incomplete trios cannot be used in this analysis without prior imputation for any missing parental genotypes (**Table 1**).

Using a TDT, the main effect of the environmental factor cannot be directly tested as all four 'individuals' (the case and three 'pseudo-controls') would be exposed to the same maternal environment. Nevertheless, the GxE effect can be estimated without the environmental main effect with the following conditional logistic regression model:

$$\text{Logit}(P[Y=1 \mid G, E]) = \beta_i + \beta_G*(G) + \beta_{GxE}*(G*E), \quad (3)$$

where Y represents the presence (1) or absence (0) of a birth defect in the infant, G represents the genotypes for the affected infant and its three pseudo-controls following an additive, dominant, or recessive model, E is the maternal environmental exposure, β_i is the model intercept which is unique to each family strata (i), β_G is the infant's genetic main effect, and β_{GxE} is the GxE effect. As an example, Beaty (2011) used the TDT in a study of 550 family trios with non-syndromic cleft palate that identified evidence of a GxE effect between maternal alcohol intake and infant genotype (in the *MLLT3* and *SMC2* genes), and between maternal smoking and infant genotype (in the *TBK1* and *ZNF236* genes).

Family-based designs such as the TDT avoid confounding due to population stratification because the analysis is conducted within families who share the same genetic ancestry (46–48). However, a major drawback is that collecting DNA samples from both parents and the affected infant can be difficult. The issue is particularly problematic for late-onset diseases since parents may no longer be alive. This concern is more easily overcome in the study of birth defects since

sample collection would be expected to occur near the time of the infant's birth when both parents are likely still alive and available to provide DNA.

Conducting a genotypic TDT analysis for GxE can be implemented using the R-package trio and the colGxE function developed by (49)), which is conveniently able to read in .ped and binary files created from PLINK software. Extensions of this approach to other family structures also exist (e.g., sibships, combinations of sibships and trios) (50–55) which can be tested using the PBAT software (32) and in R using the pbatR package (33).

Maternal versus Fetal Effects. As noted above, it may be important to evaluate GxE not only in the affected infant, but also in their mothers as well as any potential maternal-fetal GxG. For example, using data from the NBDPS, (56) found suggestive evidence that maternal genes related to metabolic conditions interacted with fetal genes related to glucose homeostasis to increase the risk of NTDs. Such interactions can be disentangled using log-linear models that adjust for maternal genetics when estimating infants' genetic effects and that adjust for infants' genetics when estimating maternal genetic effects ((42,57). This approach requires genetic information from affected and unaffected infant-mother dyads (**Figure 1**) but can also incorporate genotype information from fathers (Ainsworth et al., 2011). To evaluate GxE, we can first model main genetic effects within strata of the environmental factor then test for heterogeneity of the stratum-specific estimates. Note that a similar approach could be used for the case-control and trio designs. Hobbs et al. (2014) used an extended log-linear model to evaluate GxE effects on 616 case trios and 1645 control trios; with this approach, they found 19 maternal SNPs and 9 fetal SNPs with evidence of GxE between the genotype and maternal folate supplementation.

The log-linear approach uses a multinomial model to calculate risk ratios of genetic effects by estimating the penetrance for all possible combinations of the genotypes of the infant and the mother under a multiplicative risk model. In particular, the model estimates the effect of the infant or mother carrying one or two copies of a genetic variant on the infants risk. Then several maternal-fetal GxG (γ_{ij}) parameters given by γ_{11}, γ_{12}, γ_{21}, and γ_{22} describe the interaction between the mother's genotype and the infant's genotype in which i (j) indicates the number of copies of the risk allele that the mother (infant) possesses. For example, γ_{12} is the interaction between mother and infant genotypes when the mother possesses one copy and the infant possesses two copies.

The log-linear models must be fit within homogenous genetic populations and bias due to population stratification is possible if ancestrally varying populations are used. Implementation of this approach is easily conducted using the PREMIM and EMIM software (58). To extend this genetic model to test for GxE, the log ORs and their standard errors provided in the output of the EMIM software can be read into any standard statistical software with the ability to conduct a test of heterogeneity such as the R package, rmeta using the meta.summaries function. The test of heterogeneity is essentially testing whether the stratum-specific estimates for the genetic effect are different between the two strata of the binary environmental factor. Unfortunately, this only provides a p-value rather than a direct estimate of the GxE and is limited to binary environmental exposures but it has the advantage of discriminating between maternal GxE and infant GxE effects (**Table 1**).

Discussion

By collecting both affected and unaffected infant/mother/father triads, the NBDPS allows for flexibility in the choice of analytic approach and resulting estimates of genetic,

environmental and GxE effects. The approach used should reflect the research question of interest. For example, all described approaches can estimate the GxE effect contributed by the infant's genotype, but only the log-linear model also allows for estimation of the maternal GxE. Additionally, some causal assumptions limit the use of certain approaches. If, for instance, the genetic and environmental factors may not be independent in the study population, then the case-only approach is not appropriate. As another example, if genetic ancestry cannot be determined accurately, a genotypic TDT approach may be best since it controls for potential population stratification.

Another important point for GxE studies is how to appropriately control for potential confounding. Instead of only including the main effect of the corresponding covariate, one may also need to include interaction terms between the confounder and each of the genetic and environmental factors (Keller, 2014). The model would then control for any covariates that may confound the GxE effect, either by being correlated with the genotype or with the environmental factor. This type of adjustment is not yet common practice and so the degree of confounding due to improper adjustment is largely unknown, but it is an important consideration when assessing GxE effects (Keller, 2014).

When answering etiologic questions about interactions, an important consideration is whether the presence of interaction is assessed on the additive or multiplicative scale. All approaches presented in this paper assume a multiplicative scale, but it may be important to also consider additive interaction since the interpretation of the GxE effect, which can have public health implications, may differ based on scale (59). Some have criticized case-only designs for being only relevant to the multiplicative scale, but if certain assumptions can be made, the case-

only approach also allows inference about mechanistic interactions that are typically only identified on the additive scale (60).

Many challenges still exist when testing for GxE. Typically in GWAS, independent replication is expected for validation of observed associations. The standards for GxE studies are not as well established and thus replication of GxE associations is much less common. This may in part reflect the generally reduced power to detect interactions of GxE (in contrast with main genetic or environmental effects). Increased focus on meta-analyses could theoretically address this issue, but lack of standardization in environmental exposure definitions can make this difficult in practice. And as there are numerous combinations of genetic and environmental factors, it would be rare that the same SNP and environmental exposure are analyzed for GxE effects in multiple studies. Although for some commonly known GxE effects such as folate-related genes (i.e. *MTHFR*) and folate intake, this may be possible. Improving comparability across studies will require increased sharing of unpublished results and prior coordination among different studies to standardize collection of genetic and environmental data (61–63). Lack of standardization presents a bigger challenge if investigators wish to estimate the effect of environmental factors on a quantitative rather than a qualitative (binary) scale. For example, should smoking be measured in cigarettes per day or total number of cigarettes smoked in a narrow gestational period of relevance?

In summary, the NBDPS provides an important resource for evaluating the genetic and environmental basis of birth defects. A key component of such work is determining whether these factors work in concert to increase risk beyond their individual effects. A unique aspect of the NBDPS is the collection of samples in a manner that allow for numerous different analytic

approaches. Understanding the assumptions, strengths and weaknesses of each such approach will allow researchers to further decipher factors that increase risk of birth defects.

References

1. Etheredge AJ, Finnell RH, Carmichael SL, Lammer EJ, Zhu H, Mitchell LE, et al. Maternal and infant gene-folate interactions and the risk of neural tube defects. American Journal of Medical Genetics Part A. 2012 Oct;158A(10):2439–46.

2. Hobbs CA, Cleves MA, Karim MA, Zhao W, MacLeod SL. Maternal Folate-Related Gene Environment Interactions and Congenital Heart Defects. Obstetrics & Gynecology. 2010 Aug;116(2, Part 1):316–22.

3. Murray J. Gene/environment causes of cleft lip and/or palate: Gene/environment causes of cleft lip and palate. Clinical Genetics. 2002 May 27;61(4):248–56.

4. Sharp RR, Barrett JC. The environmental genome project: ethical, legal, and social implications. Environmental Health Perspectives. 2000;108(4):279.

5. Nelson MM, Asling CW, Evans HM. Production of multiple congenital abnormalities in young by maternal pteroylglutamic acid deficiency during gestation. J Nutr. 1952 Sep;48(1):61–79.

6. Mulinare J, Cordero JF, Erickson JD, Berry RJ. Periconceptional use of multivitamins and the occurrence of neural tube defects. JAMA. 1988 Dec 2;260(21):3141–5.

7. Bower C, Stanley FJ. Dietary folate as a risk factor for neural-tube defects: evidence from a case-control study in Western Australia. Med J Aust. 1989 Jun 5;150(11):613–9.

8. Mills JL, Rhoads GG, Simpson JL, Cunningham GC, Conley MR, Lassman MR, et al. The absence of a relation between the periconceptional use of vitamins and neural-tube defects. National Institute of Child Health and Human Development Neural Tube Defects Study Group. N Engl J Med. 1989 Aug 17;321(7):430–5.

9. Milunsky A, Jick H, Jick SS, Bruell CL, MacLaughlin DS, Rothman KJ, et al. Multivitamin/folic acid supplementation in early pregnancy reduces the prevalence of neural tube defects. JAMA. 1989 Nov 24;262(20):2847–52.

10. Laurence KM, James N, Miller MH, Tennant GB, Campbell H. Double-blind randomised controlled trial of folate treatment before conception to prevent recurrence of neural-tube defects. Br Med J (Clin Res Ed). 1981 May 9;282(6275):1509–11.

11. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. Lancet. 1991 Jul 20;338(8760):131–7.

12. Christensen B, Arbour L, Tran P, Leclerc D, Sabbaghian N, Platt R, et al. Genetic polymorphisms in methylenetetrahydrofolate reductase and methionine synthase, folate levels in red blood cells, and risk of neural tube defects. Am J Med Genet. 1999 May 21;84(2):151–7.

13. Honein MA, Paulozzi LJ, Moore CA. Family history, maternal smoking, and clubfoot: an indication of a gene-environment interaction. Am J Epidemiol. 2000 Oct 1;152(7):658–65.

14. Webber DM, Bamshad MJ, Shaw GM, Finnell RH, Shete SS, Witte JS, et al. Developments in our understanding of the genetic basis of birth defects. Birth Defects Research Part A: Clinical and Molecular Teratology. 2015;

15. Yoon PW, Rasmussen SA, Lynberg MC, Moore CA, Anderka M, Carmichael SL, et al. The National Birth Defects Prevention Study. Public health reports. 2001;116(Suppl 1):32.

16. Cogswell ME, Bitsko RH, Anderka M, Caton AR, Feldkamp ML, Sherlock SMH, et al. Control Selection and Participation in an Ongoing, Population-based, Case-Control Study of Birth Defects: The National Birth Defects Prevention Study. American Journal of Epidemiology. 2009 Oct 15;170(8):975–85.

17. Waller DK, Shaw GM, Rasmussen SA, Hobbs CA, Canfield MA, Siega-Riz A-M, et al. Prepregnancy obesity as a risk factor for structural birth defects. Arch Pediatr Adolesc Med. 2007 Aug;161(8):745–50.

18. Carmichael SL, Yang W, Feldkamp ML, Munger RG, Siega-Riz AM, Botto LD, et al. Reduced risks of neural tube defects and orofacial clefts with higher diet quality. Arch Pediatr Adolesc Med. 2012 Feb;166(2):121–6.

19. Alwan S, Reefhuis J, Rasmussen SA, Olney RS, Friedman JM, National Birth Defects Prevention Study. Use of selective serotonin-reuptake inhibitors in pregnancy and the risk of birth defects. N Engl J Med. 2007 Jun 28;356(26):2684–92.

20. Brender JD, Werler MM, Kelley KE, Vuong AM, Shinde MU, Zheng Q, et al. Nitrosatable Drug Exposure During Early Pregnancy and Neural Tube Defects in Offspring: National Birth Defects Prevention Study. American Journal of Epidemiology. 2011 Dec 1;174(11):1286–95.

21. Hobbs CA, Cleves MA, MacLeod SL, Erickson SW, Tang X, Li J, et al. Conotruncal heart defects and common variants in maternal and fetal genes in folate, homocysteine, and

transsulfuration pathways: CANDIDATE GENES FOR CONOTRUNCAL HEART DEFECTS. Birth Defects Research Part A: Clinical and Molecular Teratology. 2014 Feb;100(2):116–26.

22. Lupo PJ, Canfield MA, Chapa C, Lu W, Agopian AJ, Mitchell LE, et al. Diabetes and Obesity-Related Genes and the Risk of Neural Tube Defects in the National Birth Defects Prevention Study. American Journal of Epidemiology. 2012 Dec 15;176(12):1101–9.

23. Khoury MJ, Wacholder S. Invited Commentary: From Genome-Wide Association Studies to Gene-Environment-Wide Interaction Studies--Challenges and Opportunities. American Journal of Epidemiology. 2008 Nov 25;169(2):227–30.

24. Murcray CE, Lewinger JP, Gauderman WJ. Gene-Environment Interaction in Genome-Wide Association Studies. American Journal of Epidemiology. 2008 Nov 25;169(2):219–26.

25. Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. Two-stage designs for gene-disease association studies. Biometrics. 2002 Mar;58(1):163–70.

26. Thomas D. Gene–environment-wide association studies: emerging approaches. Nature Reviews Genetics. 2010 Mar 9;11(4):259–72.

27. Wei S, Wang L-E, McHugh MK, Han Y, Xiong M, Amos CI, et al. Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. Carcinogenesis. 2012 Aug 1;33(8):1531–7.

28. Botto LD, Yang Q. 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. Am J Epidemiol. 2000 May 1;151(9):862–77.

29. Ford RP, Tappin DM, Schluter PJ, Wild CJ. Smoking during pregnancy: how reliable are maternal self reports in New Zealand? Journal of Epidemiology & Community Health. 1997 Jun 1;51(3):246–51.

30. Newport D, Brennan P, Green P, Ilardi D, Whitfield T, Morris N, et al. Maternal depression and medication exposure during pregnancy: comparison of maternal retrospective recall to prospective documentation. BJOG: An International Journal of Obstetrics and Gynaecology. 2008 May;115(6):681–8.

31. Greenland S. Basic problems in interaction assessment. Environmental health perspectives. 1993;101(Suppl 4):59.

32. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: Tools for Family-Based Association Studies. The American Journal of Human Genetics. 2004 Feb;74(2):367–9.

33. Hoffmann T, Lange C. P2BAT: a massive parallel implementation of PBAT for genome-wide association studies in R. Bioinformatics. 2006 Dec 15;22(24):3103–5.

34. Bhattacharjee S, Chatterjee N, Han S, Wheeler W. CGEN: An R package for analysis of case-control studies in genetic epidemiology. R package version 300. 2012;

35. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 Sep;81(3):559–75.

36. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med. 1994 Jan 30;13(2):153–62.

37. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol. 1996 Aug 1;144(3):207–13.

38. Albert PS. Limitations of the Case-only Design for Identifying Gene-Environment Interactions. American Journal of Epidemiology. 2001 Oct 15;154(8):687–93.

39. Gatto NM. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. International Journal of Epidemiology. 2004 May 27;33(5):1014–24.

40. Noble EP, Jeor STS, Ritchie T, Syndulko K, Jeor SCS, Fitch RJ, et al. D2 dopamine receptor gene and cigarette smoking: A reward gene? Medical Hypotheses. 1994 Apr;42(4):257–60.

41. Chatterjee N, Kalaylioglu Z, Carroll RJ. Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, interactions and joint effects. Genetic Epidemiology. 2005 Feb;28(2):138–56.

42. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet. 2000 Jan;66(1):251–61.

43. Wang L-Y, Lee W-C. Population Stratification Bias in the Case-Only Study for Gene-Environment Interactions. American Journal of Epidemiology. 2008 May 15;168(2):197–201.

44. Neumann C, Taub MA, Younkin SG, Beaty TH, Ruczinski I, Schwender H. Analytic power and sample size calculation for the genotypic transmission/disequilibrium test in case-parent trio studies: Power and sample size calculation. Biometrical Journal. 2014 Nov;56(6):1076–92.

45. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993 Mar;52(3):506–16.

46. Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. J Natl Cancer Inst Monographs. 1999;(26):31–7.

47. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev. 2002 Jun;11(6):505–12.

48. Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol. 1999 Apr 15;149(8):693–705.

49. Schwender H, Li Q, Neumann C, Taub M, Ruczinski I. trio: Testing of SNPs and SNP Interactions in Case-Parent Trio Studies. R package version 340. 2014;

50. Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered. 2008;66(2):87–98.

51. Hoffmann TJ, Lange C, Vansteelandt S, Laird NM. Gene-environment interaction tests for dichotomous traits in trios and sibships. Genet Epidemiol. 2009 Dec;33(8):691–9.

52. Hoffmann TJ, Vansteelandt S, Lange C, Silverman EK, DeMeo DL, Laird NM. Combining disease models to test for gene-environment interaction in nuclear families. Biometrics. 2011 Dec;67(4):1260–70.

53. Lake SL, Laird NM. Tests of gene-environment interaction for case-parent triads with general environmental exposures. Ann Hum Genet. 2004 Jan;68(Pt 1):55–64.

54. Moerkerke B, Vansteelandt S, Lange C. A doubly robust test for gene-environment interaction in family-based studies of affected offspring. Biostatistics. 2010 Apr;11(2):213–25.

55. Vansteelandt S, Demeo DL, Lasky-Su J, Smoller JW, Murphy AJ, McQueen M, et al. Testing and estimating gene-environment interactions in family-based association studies. Biometrics. 2008 Jun;64(2):458–67.

56. Lupo PJ, Mitchell LE, Canfield MA, Shaw GM, Olshan AF, Finnell RH, et al. Maternal–fetal metabolic gene–gene interactions and risk of neural tube defects. Molecular Genetics and Metabolism. 2014 Jan;111(1):46–51.

57. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med. 1997 Aug 15;16(15):1731–43.

58. Howey R, Cordell HJ. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. BMC Bioinformatics. 2012;13(1):149.

59. Ottman R. Gene–environment interaction: definitions and study designs. Preventive medicine. 1996;25(6):764.

60. VanderWeele TJ, Hernández-Díaz S, Hernán MA. Case-only gene-environment interaction studies: when does association imply mechanistic interaction? Genetic Epidemiology. 2010 May;34(4):327–34.

61. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet. 2003 Mar 8;361(9360):865–72.

62. Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, et al. The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genetic Epidemiology. 2010 May;34(4):364–72.

63. Hunter DJ. Gene–environment interactions in human diseases. Nature Reviews Genetics. 2005 Apr;6(4):287–98.

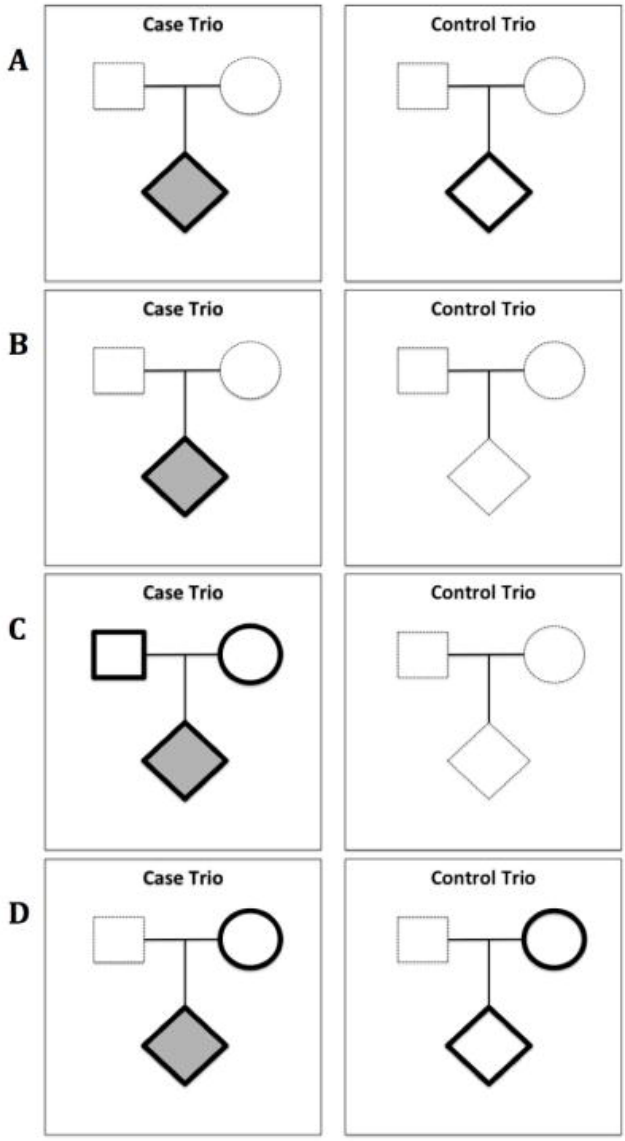
Tables & Figures

Table 2.1. Comparison of Different Gene-environment Interaction Approaches

Approach	Advantages	Disadvantages	Software Implementation	Effect Estimated
Case-control study	• Doesn't require independence of G and E	• Need to control for population stratification	• Any standard statistical software • R; `glm` function or the R-package, `CGEN` using `GxE.scan` function • `PLINK`	• Odds ratios of Infant GxE effect
Case-only study	• Improved statistical power for GxE over case-control • No controls required	• Requires independence of G and E in the source population • Cannot estimate main effect of G • Need to control for population stratification	• Standard statistical software	• Risk ratio of infants' GxE effect
Family-based trio study	• TDT unaffected by population stratification • Only affected triads required	• Missing parental genotypes prevents inclusion into study • Requires independence of G and E within families	• R package, `trio`, using colGxE function • R package, `pbatR` • PBAT	• Odds ratio of infants' GxE effect compared to pseudo-siblings (conditional on family strata)
Maternal vs. Fetal Effects study	• Log-linear model discriminates between maternal genetic effects and infant genetic effects • Can estimate maternal-infant GxG	• Genetic modeling only, cannot adjust for other covariates • Only p-value available for GxE effect • Need to control for population stratification	• EMIM software provides stratum-specific log odds ratios • GxE test of heterogeneity can be conducted in any standard statistical software for example, in R, using meta-analysis package `rmeta` using `meta.summaries` function	• Risk ratio of infant GxE interaction effect adjusted for maternal genetic effects • Risk ratio of maternal GxE interaction effect adjusted for infant genetic effects

Figure 2.1. Possible Approaches for Family Trio Data

Required data from individuals indicated by bolded black outline. Affected individuals noted by shaded shapes. Optional or not required individuals noted by dashed outline.



Panel A: Case-Control Study

Requires genotypes of unrelated affected and unaffected individuals. In a hypothetical sample of 100 case trios and 100 control trios, there would be 200 individuals in the study, 100 affected individuals and 100 unaffected individuals.

Panel B: Case-Only Study

Requires genotypes of only affected individuals. In a hypothetical sample of 100 case trios and 100 control trios, there would be 100 affected individuals in the study for this approach.

Panel C: Family-Based Trio Study

Requires genotypes of affected individuals and both their parents. In a hypothetical sample of 100 case trios and 100 control trios, there would be 300 individuals in the study for this approach, 100 affected infants and 200 parents.

Panel D: Maternal vs. Fetal Effects

Requires genotypes of affected and unaffected individuals and their mothers. In a hypothetical sample of 100 case trios and 100 control trios, there would be 400 individuals in the study, 100 affected individuals their 100 mothers, 100 unaffected individuals, and their 100 mothers.

Chapter 3

Multiple Stakeholder Views on Data Sharing in a Biobank in an Integrated Healthcare Delivery System: Implications for Biobank Governance

Introduction

Since the human genome was mapped in 2003, the past fifteen years has seen an expansion in the number of biobanks that link electronic health record data to biological samples [1]. In the U.S. alone there are over 600 biobanks [2,3] which create rich opportunities for population-based research to advance our understanding of disease mechanisms and discover new treatments and disease strategies. While we have seen significant changes in data management, high-throughput genotyping, and bioinformatics in recent years, oversight and governance mechanisms have not grown apace. The issues governing data sharing—a major goal of biobanking—require balancing difficult trade-offs between facilitating ease of sharing large amounts of genetic and clinical data to increase research and minimizing perceived and real threats to individual privacy through tight control of data access [4–6].

In 2005 Kaiser Permanente (KP) Northern California created the Research Program on Genes, Environment, and Health (RPGEH). The purpose of the RPGEH was to construct a large population-based resource to enable research to better understand genetic, behavioral and environmental factors that affect health and disease with the goal of improving the health and medical care of KP members and the general public. Participants were recruited from the adult KP membership of Northern California, which included more than three million members over the age of 18 [7]. Between 2005 and 2013, approximately 200,000 KP members volunteered to participate in the RPGEH. Participants consented to provide (a) a biospecimen (saliva and/or blood); b) access to information in their electronic medical records (EHR); and to complete c) a

survey assessing demographics, health behaviors and health history. Data available in the EHR includes a comprehensive body of clinical information such as lab test results, diagnoses, and medication use. RPGEH participants provided informed consent to share their samples and deidentified data with researchers who seek to answer a broad array of research questions. Unlike a traditional research study in which data is collected to answer predetermined research questions, a biobank such as the RPGEH collects data to answer any number of well-described research questions and operates as a resource rather than a research study. Therefore broad data collection and broad consent are essential to the RPGEH whose purpose is to be a robust resource for scientists around the globe.

From its inception, the goal of the RPGEH was to share data widely. Nonetheless, for the RPGEH, developing data access and sharing policies and procedures was complex given its situation within a health care delivery system. Beyond the traditional researcher-participant relationship, RPGEH was obliged to consider how data sharing and biobanking policies more generally impacted participants' and KP members' trust in KP. In developing the data sharing policies, the RPGEH needed to consider the potential impact on the relationship between patients and their healthcare institution; patients and their providers; as well as that between participants and researchers [8]**.** The therapeutic misconception (i.e., patients' potential confusion about the boundary between research and clinical care, and their consequent misperception that they would (or could) receive medical care as part of their research participation) may be exacerbated when a biobank is housed within a health care system. Creating a trustworthy biobank thus requires effective governance and oversight mechanisms that take into account the values and concerns of a broad range of biobank stakeholders.

Previous studies examining attitudes and values about biobank data sharing have primarily evaluated one, or at most two stakeholder perspectives [9]. Typically studies have included biobank participants (or potential participants such as the general public) or study investigators [10–12]. Some studies have included other governance experts such as biobank managers or administrators, but none of these studies have been within the context of a health care delivery system [10,13,14]. In this paper we present results from in depth qualitative interviews with multiple stakeholder groups in order to understand their perspectives about wide data sharing in the RPGEH. The interviews were conducted between 2010-2011, during a time when the collaboration and access policies and an Access Review Committee (ARC) were being developed but not yet finalized. In addition, how the RPGEH would share data with the NIH Database of Genotypes and Phenotypes (dbGaP) was being discussed.

We sought to answer the following question in order to help guide RPGEH data access policies and procedures: What are different stakeholders' values and preferences regarding protection of participants' privacy and sharing participant data? Our findings address three main topics: (1) perceived benefits and harms of data sharing, specifically perspectives on individual and social harms associated with data sharing; (2) governing access to RPGEH data, particularly the perspectives on management and governance of RPGEH data; and (3) the blurred boundary between research and clinical care in the context of biobanking.

Methods

Study Design

We conducted 46 semi-structured interviews, in person or via phone, and two focus groups with seven stakeholder groups:

1. RPGEH participants: KP members who consented to participate in the RPGEH.

2. RGPEH decliners: KP members who were contacted and actively declined participation in RPGEH.
3. Kaiser Permanente Research scientists: PhD or MD-level scientists working within the Division of Research at Kaiser Permanente.
4. External scientists: Scientists from neighboring academic institutions, who would likely use the resource.
5. KP leadership: Members of the RPGEH executive oversight committee.
6. KP Human Subjects Research Protection Program staff (HRPP): Staff responsible for reviewing and ensuring human subjects protections at KP.
7. RPGEH Community Advisory Panel (CAP) members: Community stakeholders who advised the RPGEH program on policies, procedures, and community outreach.

Community Advisory Panel members and RPGEH participants and decliners received a $35 gift card for their participation. This study was approved by the Kaiser Foundation Research Institute Institutional Review Board (IRB).

Sample Description

Interview and focus group participants were selected using a purposive sampling strategy to identify key informants from each stakeholder groups. 45 individuals participated in an interview and seven CAP members participated in one of two focus groups. Numbers are shown for each stakeholder group in **Table 1**.

Data Collection and Analysis

We developed an interview and focus group guide and pilot tested each with multiple stakeholders. Questions addressed a wide range of domains including informed consent, data sharing, benefit sharing, community harms, motivations/barriers to participation and depositing

of data in dbGaP among others. All interviews and focus groups were audio-recorded and transcribed.

We employed a constant comparative approach in which the analysts discussed and identified themes that emerged from the codes [15]. All coding was conducted using ATLAS.ti version 7 (ATLAS.ti Scientific Software Development GmbH, Berlin, Germany) Three social scientists (JHW, PL, CPS) conducted an initial review of all transcripts and developed analytic memos identifying major topics emanating from each stakeholder group. They developed a codebook through a reflexive, iterative process. The analysts coded a test set of transcripts to ensure high intercoder reliability. Any discrepancies between coders were discussed and a decision was made by the team on which code to apply. After sufficient training, at least two analysts coded the transcripts with the codebook but analysts also conducted open coding to identify new themes, subthemes, and phenomena. Finally, a fourth analyst (CT) analyzed the coded transcripts, focusing on the themes that emerged across all stakeholder groups around (1) benefits and harms of data sharing, (2) appropriate balance between protecting individual privacy and wide data sharing; and (3) concerns with establishing a biobank within a healthcare system, such as blurred boundaries between clinical care and research

Results

Perceived Benefits and Harms of Data Sharing

Stakeholders all discussed the importance of sharing data so that it benefitted society. For example, stakeholders believed sharing the data could contribute to the development of new scientific discoveries or improve our understanding about certain diseases especially those that have affected their lives personally such as having a relative with cancer.

Benefit: Moral Obligation

Most KP scientists felt that RPGEH data should be shared as a moral obligation. Many scientists explained that RPGEH data is a valuable and unique resource, and most believed that not allowing access to qualified scientists would be an ethical violation.

> [KP Scientist 109] *"... it's a tension…You create this big resource, and it's unethical not to try to make the most of it. And so, what do you do about that?"*

Other stakeholders also brought up the moral obligation to share data with the NIH which funds genome-wide association studies being conducted with RPGEH data.

> [External Scientist 112] *"This is a huge investment from the National Cancer Institute--$35 million, is my understanding. So that's taxpayer money. That is—so, the citizens of the United States are paying Kaiser (a private corporation) to produce this resource."*

> [KP leader 187] *"I'm also aware that the funding agencies feel that if federal dollars are paying for it, they want to have something that is then generally available."*

Perceived harms from data sharing, and from establishing the RPGEH in general, fell into two main categories, individual harms and social harms.

Individual Harm: Re-identification

The collection of genetic information creates an uncertain though definite risk in which there is a possibility that participants could be re-identified even after traditional unique identifiers (i.e. name, address, etc.) have been removed due to the uniqueness of each person's set of genetic information.

> [HRPP staff 106] *"I think it's raised the topic of how confidential is a person's genome...it ultimately becomes a unique identifier, which, if their genetic analysis was also in some other database, things might be able to be cross-referenced"*

This concern was repeated by participants who were also concerned about the identifiability of genetic data.

> [Participant 142] *"The DNA, the genetic makeup of any person, is so unique that no one has that. You know?"*

Individual Harm: Use of Research Data for Medical Underwriting

Many KP members understand KP as one large organization, in which all entities within KP are interconnected. They view their insurance coverage as being governed by and controlled by the same institution that provides their medical care. Given this perspective, many participants, decliners, CAP members, and one HRPP staff feared that RPGEH data, specifically the genetic information, could be used for medical underwriting, a practice in which a person's medical information is used to evaluate their application for medical coverage.

> [Decliner 182] *"It could be misused by Kaiser. It could be misused by governmental boards that are going to review who gets care and who doesn't get care."*

> [Participant 172] *"So, you know, you look at the results and say, 'Well, we can't help this person, at least not in the immediate future, and so we're not going to pay for any treatment.' I mean, that's really the downside and the, you know, the tradeoff the insurance companies make."*
>
> [CAP Member A] *"So I think the biggest challenge is: How will they use that data in terms of avoiding the practice of medical underwriting? Because healthy patients are cheaper to take care of than people who are ill or may became sicker. So I think that's why a lot of people could be uneasy about it."*

[HRPP staff 106] *"And then, yeah, even within Kaiser, could it somehow be used for underwriting purposes? You know, people are more at risk and therefore gradually charged higher premiums, lower premiums, family members—I don't know. But that's the main risk I would see about the integrated delivery system and the special risks that we have"*

This perceived fear is unlikely a true possible harm since Genetic Information Nondiscrimination Act (GINA), passed in 2008, provides protections against medical insurance discrimination on the basis of genetic information.

Social Harm: Stigmatization of Groups Based on Research Findings

Many KP scientists expressed concerns that external scientists' unfamiliarity with KP data may cause them to use the data improperly, either by evaluating inappropriate research questions or by misinterpreting the meaning of data fields/variables. RPGEH participants, decliners and scientists believed that KP's reputation could be tarnished if any scientist used KP data to reinforce harmful stereotypes. Even if the original intent was not malicious, they believed that data could be misinterpreted or incorrectly analyzed due to a lack of understanding about KP's clinical data or demographic distribution of its members.

[Decliner 155] *"My only concern is how it would affect you guys, Kaiser, if you would get blamed for seemingly supporting a certain stereotype or adding on to a certain stereotype, and how people, you know, can misconstrue things"*

[KP Scientist 105] *"I don't see data sharing, per se, being a problem. [But] I think if data are poorly used or misused, then that could come back to reflect on us."*

Even if there was no misuse of data, there is still a possibility that certain populations can experience harm. For example, if data was used to find higher prevalence of a disease in a certain subgroup, there could be unanticipated harms from such a result.

> [Decliner 169] *"…if you basically say that your research can prove that there's a certain racial group that basically has a high chance of getting certain diseases, you know, and then say, 'Well, because of this research, maybe you either want to charge them a higher premium for their insurance, or we're just not going to cover them.'"*

Community Advisory Panel members were concerned about social harm from research that could stigmatize certain populations and suggested several possible mechanisms to minimize this potential:

> [CAP Member B] *"I'm wondering. I think it's an idea; can there be a social responsibility section* [of the proposal to obtain access to RPGEH data] *where the research would have to explain how they're going to use the outcome of the research. I don't know how that would work because I've never seen one but it's something."*

Another CAP member suggested that an additional layer of review beyond that of an IRB be included to provide community input for studies that dealt with marginalized populations especially if they pertained to stigmatized conditions such as alcoholism.

> [CAP Member C] *"So, I mean, even though there are major issues with colon cancer and not getting insurance, I think alcoholism has, like, a larger social stigma with it and connections with, you know, law, etc., that could be problematic... I think there should be an additional review board in terms of community members overlooking, you know, research study, progress, and then, you know, paper that they're going to write."*

Social Harm: Loss of Trust

The research enterprise rests heavily on volunteer participants. If individuals lose trust in scientists or the research establishment, they would likely not participate in research and scientific progress may be stymied. One reason why individuals may stop participating in research is if they lose faith in researchers as responsible stewards of their data. Being a responsible steward would include ensuring that data protections are in place. One KP leader described how patient privacy is of paramount importance over all else because this loss of trust would put the entire research program in jeopardy.

> [KP Leader 188] *"…if we don't protect the privacy of our patients, they will make us stop doing this and the greater good will be harmed."*

Kaiser Permanente has garnered a reputation among its members as a trustworthy organization. This is likely due to their positive previous experiences with KP as a medical provider. Many participants in RPGEH are long-term members of KP (over half had been KP members for 20 years or more).

> [Participant 142] *"I trust Kaiser. I've been with Kaiser all my life. You know, when I got married, I continued with Kaiser. And I've never had problems with Kaiser."*
>
> [Participant 148] *"…once I found out that it [RPGEH] was through Kaiser, who I've been with since I was a little kid, then I was all for helping"*

The trust in KP is firmly rooted in the relationship between patients and their providers. Once this foundation is built, patients are willing to take the next step with KP into being research participants.

> [Participant 144] *"…actually, this [RPGEH] makes me feel more positive about Kaiser. I mean, first of all, understand, I was born at Kaiser. So that's the only health care I've ever known. And my children were born there…And it makes me feel more warm and*

fuzzy to know that they're engaged in cutting-edge research and that, you know, just us average people can make some kind of contribution and participate. So, bring it on!"

Many stakeholders were aware that trust in KP is a common sentiment among its members and is central to KP scientists' ability to continue conducting research at KP.

[KP Leader 187] *"And finally, the loyalty of our members, which results in our ability to follow them over decades and sometimes through generations… that we continue to enjoy that confidence and trust of our members in us to do this kind of important research…And we need to, first of all, assure ourselves that there are the appropriate firewalls and training and agreements and monitoring that can assure ourselves as well as our employees and our members that we are, again, good custodians of their trust."*

Many stakeholders were cautious to mention that sharing data from a resource like RPGEH could have negative consequences to its reputation, especially regarding data security and respect of RPGEH participants' privacy.

[KP Leader 188] *"I think that what became clearer is that this was potentially an extremely large endeavor and one of potentially great impact and also one of – that carried unique kinds of risks to reputation, in particular – reputation and concerns about privacy in particular."*

[KP Scientist 101] *"I think the only way it could really impact us negatively is if, number one, there's real harms that occur to patients. There is a breach in confidentiality, and some harms do happen to patients, which I think is extremely unlikely, especially both of those pieces."*

Governing Access to Data

One aspect of data governance is the development of data access policies. Every biobank has its own requirements for determining whether a scientist or entity should be granted data access. Criteria for evaluation could include elements such as research goals, plans for data protection, or requester's affiliation. Stakeholders were asked what should be considered when granting access to RPGEH data.

Scientist's Affiliation

All stakeholders agreed that creating RPGEH and sharing data had the potential for broad benefit in terms of understanding disease and health. But there were some differences amongst stakeholder groups about how data should be shared and with whom. Many participants and decliners in particular had concerns with scientists affiliated with a for-profit institution such as a pharmaceutical or biotech company but had fewer concerns with academic scientists. Largely, stakeholders did not object to data access by external scientists as long as they were not affiliated with for-profit entities.

> [Decliner 155] *"I think that you should definitely accept other medical researchers, such as yourselves…You mentioned, like, biotech companies earlier. I was kind of surprised; I wasn't expecting that type of organization, that kind of company, as being involved."*
>
> [Participant 137] *"I guess the corporate people who are trying to benefit from, I don't know, making profit off of it, in the sense that it's not doing anything for the greater good. That would be my only issue with it."*
>
> [Participant 145] *"I kind of felt that it was more geared towards, like, you know, educational, like schools and things like that. Not really, like, something like biotech or like Genentech and things like that"*

Although academic affiliations seemed preferable, some participants thought that it would still be beneficial to share data with private entities, such as for-profit pharmaceutical companies, who might be able to translate discoveries into applicable therapies.

> [Participant 144] *"I'm more inclined to favor public institutions, you know. But on the other hand, I think it's the privates that can translate pure research into something—some technique or pill or whatever—that is actually beneficial, that can be applied to individuals who need therapies. So I believe that public-private partnerships are worthwhile."*

In addition to concerns about for-profit institutions, several RPGEH participants voiced concerns about non-US scientists and foreign governments accessing RPGEH data.

> [Participant 141] *"Well, I guess I would strongly like it to stay within the United States and not go outside of that for research, because I'm not sure what other countries—I know that they have positive ends of their research, but I know for some others it may not be a positive thing."*
>
> [Participant 139] *"I would be wary about foreign governments and terrorists somehow having access to the data."*

Loss of Control Over Data

A fear of government was a recurring topic and was usually expressed in terms of multiple perceived risks including forensic use of biobank data, vulnerability to political agendas, and a general loss of control over the data. RPGEH participants and CAP members explained that government control of the data would allow it to be used for forensic purposes and lead to violations of individuals' privacy.

[Participant 142] *"And I also don't want it so that someone could break into the database there and say, "Oh, hey! We've got a match for a murderer."*

[CAP Member A] *"There are other DNA databases out there, like criminal DNA databases…So I think that would be a concern also. But I think, for me, the question would be who would have access to it."*

Stakeholders pointed out that the NIH may be less trustworthy than KP because a government-sponsored agency could not remain purely objective or uninfluenced due to the influence of politics.

[Participant 172] *"…because NIH is a federal political organization as well as a government organization, is that politics will get involved…And so they basically are very sensitive to political winds, and I know that the NIH tries hard to at least, personally, I would try very hard to avoid that—but being a political bureaucracy with public funding…they are going to have political masters"*

[CAP Member C] *"…the NIH is controlled by Congress and politicians. It's not controlled by somebody who's saying, 'Oh, let's go talk to the community.'"*

KP leadership and HRPP staff acknowledged the benefits of sharing externally but they were concerned about protecting the individual privacy of KP members and that it would be more difficult to do so once data is shared outside of KP. They believed that external scientists who may not have a direct connection to biobank participants may not handle RPGEH data with the same care as internal KP scientists.

[KP Leader 187] *"It's essential that we do collaborate, but how can we collaborate in a manner that continues to make our members volunteer and the people who have volunteered comfortable that we're being good custodians of their data and their DNA?"*

[HRPP 111] *"In a sense, we're trying to be the parent, if you will, of this database in saying, 'Well, you can interview my child'—I'll use this analogy just a little bit—'You can interview my child, but really, there are some questions that you can ask and some questions that you can't ask.' You know, we have to take some responsibility for, in a sense, the behavior of the researcher."*

Most stakeholders were concerned that KP would be required to relinquish its control over access to RPGEH data when it deposited in dbGaP because scientists who apply to use dbGaP data are evaluated by a NIH data access committee, whose members have no ties to KP or KP members.

[KP Scientist 107] "*The problem with dbGaP is that, to the extent that it's a problem or even a perceived problem, is that here's a database that ultimately is not controlled by you, and it's not controlled by a person of your choosing."*

Blurred Boundary Between Research and Clinical Care

There is a long history in research ethics of defining a clear boundary between clinical care and research. [16] This stems from of a concern—referred to as the "therapeutic misconception"—that patients who participate in research may assume that their participation, especially one coordinated by their healthcare provider, will lead to personal health benefits. To minimize this confusion, which could be considered a source of coercion to participate, the RPGEH consent form states that "This research is not intended to benefit individual participants. However, we hope that results from the research will improve how doctors prevent, diagnose and treat major illness in the future. You will not receive personal health or medical results from taking part in the RPGEH."

The potential for therapeutic misconception may be greater in an integrated delivery system like KP since members receive all services from a single health system and are accustomed to their data being linked. Some RPGEH decliners believed that by participating in research, their data may reveal an underlying disease leading them to undergo treatment for a condition that might be asymptomatic or non-life threatening:

> [Decliner 154] *"I thought that you would find something that I didn't know I had. And, you know, then I'd be in another whole ball game, you know? Now I'd be in treatment, and all I was trying to do was, you know, be helpful. Yeah, I didn't want to."*

Human Research Participant Protection Program staff were worried that RPGEH participants would not be able to distinguish a boundary between research and clinical care, and that this would have negative consequences.

> [HRPP staff 106] *"And, of course, for members it might be somewhat confusing, because they think they're going to get something out of it, even though consent forms generally probably tell them they aren't, that they're not going to actually benefit from participation. But they night somehow think they might."*

Conversely, several scientists expressed that the blurring of the boundary between research and clinical care could be beneficial. They explained that new opportunities for data linkage between clinical, environmental, and genetic data, RPGEH could position KP to become more of a pioneer in translational research, bringing scientific discoveries into the clinical setting. Some KP scientists believed this is especially true in the setting of KP where researchers have opportunities to interact with clinicians who want to learn more about their patient population and how to better treat them.

[KP Scientist 107] *"Having the research done here gives them [KP clinicians] an inside track on what might turn out to be important diagnostically, therapeutically."*

[KP Scientist 108] *"I think it will promote the translation of our findings into clinical practice, because it's in the context of this, and we've got a big organizational machinery that's looking for that. And think we'll help take the research findings and move them into the clinical realm."*

[KP Scientist 108] *"And the other really big important advantage is that I can, you know, collaborate with people internally on projects that I wouldn't normally be—it's hard to collaborate, you know."*

Discussion

Unsurprisingly, a key finding from our study was that there was broad support across stakeholder groups for creating a shared research resource. Given that this was a common theme found across multiple research studies on biobanking over the past decade [9,17–21], we focused our analyses on stakeholders' perspectives regarding the risks and concerns of sharing data, attitudes towards governance of the data, and issues around the blurry boundary between research and clinical care.

Stakeholders identified multiple individual harms including loss of privacy through genetic re-identification, medical underwriting, and forensic use of genetic biobank data by government entities. These individual harms were consistent with other qualitative and survey-based research studies conducted on public and patient attitudes towards biobanking [8,9,17,19–23]. While medical underwriting and insurance discrimination were cited by participants and decliners, these perceived risks were cited by HRPP members as well. This was an interesting finding since it is not consistent with the way in which KP separates its clinical activities from its

research activities and its underwriting responsibilities. A previous study comparing the views of genetic researchers and IRB professionals also found that individuals involved in the ethical review of research are more likely than researchers to believe that insurance discrimination is a real threat [24]. There is federal legislation that helps to mitigate some individual harms and the RPGEH communicated these protections (as well as the limitations of these protections) to participants during the consent process. For example, to address many individual harms, GINA and Certificates of Confidentiality provide legal protections to protect individuals' rights. This suggests that the fear of health insurance discrimination may not be easily addressed by assuring participants and other stakeholders that there are regulations in place to protect individuals. Understanding the underlying fears and the larger context may lead to more appropriate methods for addressing stakeholders' deeply held concerns in this area, this seems especially necessary in the recruitment of non-white participants for biobank research [22,23,25].

The concerns expressed around social harms pertained to the view that findings from studies could stigmatize certain groups, and also that stigmatization could lead to loss of patient trust and decreased rates of research participation. Understanding social harms in the context of biobanking is particularly important given that the individual harms such as loss of privacy and insurance discrimination are generally well monitored and addressed by IRBs but social harms are typically beyond the scope of an IRB review. Far fewer studies have examined stakeholders' perceptions of social, compared to individual, harms from participating in biobanks. One previous study of cancer patients expressed concerns that using genetic information in research could cause stigmatization of certain racial or ethnic groups [22]. While there are very few studies investigating the views of scientists, leaders, or HRPP members about social harms, these concerns are consistent with other studies of biobank participants and decliners [8,18]. One

concern identified mainly by scientist stakeholders and KP leadership was that the RPGEH Access Review Committee does not directly control data access when data are shared through dbGaP. Although this is true, sharing RPGEH data through dbGaP is highly prescribed and specific restrictions are in place through a data use certification agreement that minimizes the likelihood of stigmatization of certain groups [26].

When probed about whether a scientist's affiliation should be taken into account when governing data access participants and decliners most often cited fear of for-profit entities, government agencies, and non-US scientists. Sharing data with the NIH, a federal agency, through dbGaP also concerned many other stakeholders including scientists, KP leaders, and CAP members. However, the concerns of scientists and biobank leaders was that control over RPGEH data would no longer be in the hands of KP employees; while participants, decliners, and CAP members were afraid of law enforcement agencies using RPGEH data for forensic purposes. Concerns about non-US scientists in our study has been echoed by studies of public opinion in other countries [18,27,28] which cite exploitation and lack of reciprocal benefit to research participants as the main reason for skepticism of foreign scientists. Many stakeholders were concerned that for-profit sponsored research could lead to additional group harms such as profiteering resulting in research that would not improve societal health outcomes (i.e. because therapeutics developed from the research may be cost prohibitive for people in the biobank).

Finally, one of the more interesting findings from our research related to the complicated issue of the blurred boundary between research and clinical care at KP. All participants and decliners we interviewed perceived KP, their care provider, as the same entity as KP, the research institution. While a blurry research-clinical interface can be seen as an individual harm that may constitute a therapeutic misconception it also can be viewed as a social benefit since

research findings may be more quickly translated into clinical practice or new therapeutics [29]. In addition patients may be more likely to trust KP to conduct research with their information because of their existing trust in the KP organization. Other studies have also shown a relationship between trust in a healthcare system and research participation where having a care provider whom they actively visit is a strong predictor of agreeing to participate in research involving wide data sharing [8,30,31].

The main objective of this study was to help inform data access policies for the RPGEH. In part as a result of our research findings, along with feedback from RPGEH leadership, the RPGEH created a robust Access Review Committee (ARC) that emphasized the importance of rigorous evaluation of data access requests. The goal of this two-step interdisciplinary ARC was to serve as a major data governance mechanism to maintain participant trust, and minimize individual and social harms through fair, consistent, and transparent access policies and procedures. ARC members include KP Division of Research investigators and biostatisticians, including an investigator with expertise in ethical, legal, and social issues (ELSI); external stakeholders, and additional investigators as required by the projects under consideration. Proposals to obtain RPGEH data are reviewed for multiple factors, including scientific merit; the scientific qualifications of the applicant; appropriateness of RPGEH as a resource for the proposed research; and feasibility (given the RPGEH resources). In order to address potential social harms, the application also includes a specific question related to ELSI asking the applicant to discuss any potential individual or social/group harms that might occur as a result of conducting the research, interpreting or disseminating research findings and to suggest strategies to mitigate those harms. Additionally, a community member was also added to the ARC to contribute to this effort and to maintain community representation in this process [32] .

Currently, the RPGEH is being integrated into a larger KP national biobank (KP Research Bank), which similarly has an ARC that serves the same purpose and conducts the same activities as the RPGEH ARC.

This study has some limitations. The study represents a description of opinions collected from a relatively small sample of participants from each stakeholder group. The interviews were conducted at an early point in the creation of the RPGEH, during the process of infrastructure building and policy development but before the collection of genetic data. As such, it depicts concerns and opinions about the operation of a biobank as it was being built. It should also be noted that the stakeholders interviewed were largely KP members and/or individuals either employed by KP or who had experience collaborating with KP. Patients in other health care delivery systems or scientists without experience collaborating with KP may have voiced different concerns. Despite these limitations, our findings provide useful information for the institutions, particularly health care organizations, developing other biobanks about the concerns of diverse stakeholders, especially those related to the potential for social harms.

References

1 Shoenbill K, Fost N, Tachinardi U, Mendonca EA: Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. J Am Med Inform Assoc 2014;21:171–180.

2 Boyer GJ, Whipple W, Cadigan RJ, Henderson GE: Biobanks in the United States: How to Identify an Undefined and Rapidly Evolving Population. Biopreservation Biobanking 2012;10:511–517.

3 Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, et al.: Characterizing biobank organizations in the U.S.: results from a national survey. Genome Med 2013;5:3.

4 Kaye J, Briceño Moraia L, Mitchell C, Bell J, Bovenberg JA, Tassé A-M, et al.: Access Governance for Biobanks: The Case of the BioSHaRE-EU Cohorts. Biopreservation Biobanking 2016;14:201–206.

5 Harris JR, Burton P, Knoppers BM, Lindpaintner K, Bledsoe M, Brookes AJ, et al.: Toward a roadmap in global biobanking for health. Eur J Hum Genet 2012;20:1105–1111.

6 O'Doherty KC, Burgess MM, Edwards K, Gallagher RP, Hawkins AK, Kaye J, et al.: From consent to institutions: Designing adaptive governance for genomic biobanks. Soc Sci Med 2011;73:367–374.

7 Gordon N: Characteristics of Adult Health Plan Members in Kaiser Permanente's Northern California Region, as Estimated from the 2011 Member Health Survey. Oakland, CA, Division of Research, Kaiser Permanente Medical Care Program, 2013.Available from: http://www.dor.kaiser.org/external/mhs11reg

8 Beskow LM, Burke W: Offering individual genetic research results: context matters. Sci Transl Med 2010;2:38cm20.

9 Garrison NA, Sathe NA, Antommaria AHM, Holm IA, Sanderson SC, Smith ME, et al.: A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. Genet Med 2016;18:663–671.

10 Husedzinovic A, Ose D, Schickhardt C, Fröhling S, Winkler EC: Stakeholders' perspectives on biobank-based genomic research: systematic review of the literature. Eur J Hum Genet 2015;23:1607–1614.

11 Oushy MH, Palacios R, Holden AEC, Ramirez AG, Gallion KJ, O'Connell MA: To Share or Not to Share? A Survey of Biomedical Researchers in the U.S. Southwest, an Ethnically Diverse Region. PLOS ONE 2015;10:e0138239.

12 Rivera SM, Goldenberg A, Rosenthal B, Aungst H, Maschke KJ, Rothwell E, et al.: Investigator Experiences and Attitudes About Research With Biospecimens. J Empir Res Hum Res Ethics 2015;10:449–456.

13 Moodley K, Singh S: "It's all about trust": reflections of researchers on the complexity and controversy surrounding biobanking in South Africa. BMC Med Ethics 2016;17. DOI: 10.1186/s12910-016-0140-2

14 Colledge F, Elger B: Getting a Fair Share: Attitudes and Perceptions of Biobank Stakeholders Concerning the Fairness of Sample Sharing: Getting a Fair Share: Attitudes and Perceptions of Biobank Stakeholders. Bioethics 2015;29:424–430.

15 Glaser BG, Strauss AL: The discovery of grounded theory: strategies for qualitative research. 4. paperback printing New Brunswick, Aldine, 2009.

16 Kimmelman J: The Therapeutic Misconception at 25: Treatment, Research, and Confusion. Hastings Cent Rep 2007;37:36–42.

17 Rahm AK, Wrenn M, Carroll NM, Feigelson HS: Biobanking for research: a survey of patient population attitudes and understanding. J Community Genet 2013;4:445–450.

18 Nicol D, Critchley C, McWhirter R, Whitton T: Understanding public reactions to commercialization of biobanks and use of biobank resources. Soc Sci Med 2016;162:79–87.

19 Trinidad SB, Fullerton SM, Bares JM, Jarvik GP, Larson EB, Burke W: Genomic research and wide data sharing: Views of prospective participants. Genet Med 2010;12:486–495.

20 Rogith D, Yusuf RA, Hovick SR, Peterson SK, Burton-Chase AM, Li Y, et al.: Attitudes regarding privacy of genomic information in personalized cancer therapy. J Am Med Inform Assoc 2014;21:e320–e325.

21 Lemke AA, Wolf WA, Hebert-Beirne J, Smith ME: Public and Biobank Participant Attitudes toward Genetic Research Participation and Data Sharing. Public Health Genomics 2010;13:368–377.

22 Pentz RD, Billot L, Wendler D: Research on stored biological samples: Views of African American and White American cancer patients. Am J Med Genet A 2006;140A:733–739.

23 Lemke AA, Halverson C, Ross LF: Biobank participation and returning research results: Perspectives from a deliberative engagement in South Side Chicago. Am J Med Genet A 2012;158A:1029–1037.

24 Edwards KL, Lemke AA, Trinidad SB, Lewis SM, Starks H, Snapinn KW, et al.: Genetics researchers' and IRB professionals' attitudes toward genetic research review: a comparative analysis. Genet Med 2012;14:236–242.

25 Heredia NI, Krasny S, Strong LL, Von Hatten L, Nguyen L, Reininger BM, et al.: Community Perceptions of Biobanking Participation: A Qualitative Study among Mexican-Americans in Three Texas Cities. Public Health Genomics 2017;20:46–57.

26 dbGaP Data Use Certification AgreementAvailable from: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000674.v2.p2

27 Chen H, Gottweis H, Starkbaum J: Public Perceptions of Biobanks in China: A Focus Group Study. Biopreservation Biobanking 2013;11:267–271.

28 de Vries J, Munung SN, Matimba A, McCurdy S, Ouwe Missi Oukem-Boyer O, Staunton C, et al.: Regulation of genomic and biobanking research in Africa: a content analysis of ethics guidelines, policies and procedures from 22 African countries. BMC Med Ethics 2017;18. DOI: 10.1186/s12910-016-0165-6

29 Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL: An Ethics Framework for a Learning Health Care System: *A Departure from Traditional Research Ethics and Clinical Ethics*. Hastings Cent Rep 2013;43:S16–S27.

30 Brothers KB, Morrison DR, Clayton EW: Two large-scale surveys on community attitudes toward an opt-out biobank. Am J Med Genet A 2011;155:2982–2990.

31 Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AHM, Aufox SA, et al.: Public Attitudes toward Consent and Data Sharing in Biobank Research: A Large Multi-site Experimental Survey in the US. Am J Hum Genet 2017;100:414–427.

32 Haldeman KM, Cadigan RJ, Davis A, Goldenberg A, Henderson GE, Lassiter D, et al.: Community Engagement in US Biobanking: Multiplicity of Meaning and Method. Public Health Genomics 2014;17:84–94.

Table 3.1. Interviews/focus groups by stakeholder group

Stakeholder Group	N	Method
Scientists		
KP Scientists	10	Individual Interviews
External Scientists	5	Individual Interviews
HRPP Staff	4	Individual Interviews
KP Leadership	3	Individual Interviews
CAP Members	7	Focus groups (2)
KP Members Eligible for RPGEH		
RPGEH Participants	14	Individual Interviews
RPGEH Decliners	9	Individual Interviews
Total Participants	52	

Funding

The work conducted in this dissertation was supported by the following funding sources.

National Cancer Institute grants: F31 Pre-doctoral Fellowship Grant F31CA200139, PI: Tai and R01 Grant R01CA088164, PI: Witte.

University of California, San Francisco awards: Earle C. Anthony Social Sciences Fellowship Award and Matilda Edlund Merit Scholarship Award.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date 6/8/18