Importance, Challenges, and Opportunities of Gene-Environment Interactions (GxE) Research: A Study of Parkinson's Disease

Nirupama Nini Shridhar

A dissertation

submitted in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Karen Edwards, Chair

Erika Blacksher

Cyrus Zabetian

Yvonne Lin

Karl Hill

Program Authorized to Offer Degree

Public Health Genetics

University of Washington

Abstract

Importance, Challenges, and Opportunities of Gene-Environment Interactions (GxE) Research: A Study of Parkinson's Disease

Nirupama Nini Shridhar

Chair of the Supervisory Committee

Dr. Karen Edwards

Department of Epidemiology

Objective. This interdisciplinary dissertation comprised a normative analysis: social justice in genomics research, and a statistical and functional analyses: evaluating gene-environment interactions (GxE) in Parkinson's disease (PD) research. The normative analysis examined the potential for unfair health distribution created by the genomics research agenda, and evaluated research methodologies that may be more socially just. The statistical and functional analyses evaluated gene-environment interactions GxE in PD. GxE are thought to be causal in complex disease risk. Understanding the mechanisms by which GxE alter risk in complex diseases represents a leverage point by which to modify disease biology. The intent here, was to find new susceptibility genes that are mutable (by environmental exposures), and in the causal pathway to PD risk. Smoking has been consistently shown to be associated with reduced risk of developing PD. One of the biggest effects of smoking is DNA methylation. This research used a

novel gene selection method by following the environment into the body, by identifying genes that are known to be methylated by smoking. This set up an *a priori* hypothesis that somewhere along the network of genes known to be methylated by smoking is an interaction with gene networks involved in PD risk. The research also hypothesized that regulatory single nucleotide polymorphisms (rSNPs) on genes that are differentially methylated by smoking act as genetic determinants to epigenetic modifications.

Methods. Normative Methods: A normative bioethical analysis was conducted in looking at the question of "what would be a socially just genomics research agenda?''. The analysis used Sen's 'capabilities approach', and work by other capabilities theorists, to define the framework for social justice. This framework was then applied to genomics research to examine social justice in genomics research.

Statistical and Functional Methods: GxE association analysis were performed using case-control data from the NeuroGenetics Research Consortium (NGRC). The NGRC data contain 2000 PD cases and 1986 unrelated controls (n= 3986). The dataset, included genotype information for the SNPs on the 39 genes selected (n= 2281), and the environmental factor (smoking use). The genes were tiered based on evidence: Tier 1 (3 genes and 747 SNPs), and Tier 2 (36 genes and 1534 SNPs). The threshold for statistical significance (after adjusting for multiple comparisons) were: Tiers 1 and 2 ($p=1.13 \times 10^{-4}$ and $p=5.67 \times 10^{-5}$). Based on the results of the association analysis, the SNP results were then further characterized for functional consequence by synthesizing multiple lines of evidence to identify PD susceptibility genes from the perspective of the environmental exposure.

Results. Normative Analysis: Gene-centric approaches do not meet the demands of social

justice. GxE approaches have the potential to meet the demands of social justice. It is a socially just method because it could to lead to translational outcomes with public health utility. Association and Functional Analyses: No interactions were statistically significant (in either tier). They were however, statistically suggestive. Layering the functional evidence over the statistical evidence, a single gene emerged in Tier 2 that was statistically suggestive and with significant functional evidence. *RARA,* a gene in the retinoic acid signaling pathway coding for the protein RARα. The two SNPs showing evidence for functional consequences, (rs2120200 and rs36030243) are SNPs that alter gene expression and bind transcription factors. Results from the association analysis also had two SNPs from the same gene among the top 10 SNPs in the sensitivity analysis (rs12103711 and rs2715553). Rs12103711 is an intronic SNP of *RARA* (OR = 1.80, 95% CI: 1.23- 2.65, p = 0.002802) and rs2715553 is a cSNP of *RARA* (OR = 0.72, 95% CI: 0.57 - 0.91, p = 0.00563). Linkage disequilibrium(LD) was evaluated between the four SNPs in *RARA* (two from the functional analysis, and two from the statistical analysis) to determine if these SNPs were all pinpointing a single signal with evidence for regulation, and found that the four SNPs are in strong LD with each other, although the rs2715553, a cSNP is physically some distance away (35 kb) from the other three SNPs. We found rs2120200 and rs36030243 are likely rSNPs on the *RARA* gene that interact with smoking and are in LD with a coding SNP on the same gene.

Conclusion. To our knowledge, no prior analysis has used a capabilities approach to examine social justice in genomics research. We developed a framework by which to analyze social justice in genomics research, and using a case study of Parkinson's disease (PD), we demonstrated that gene-environment interaction (GxE) methods meet the demands of social

justice in genomics research. This exploratory analysis demonstrated that selecting genes by following the effect of the environmental exposure is a valid method of identifying susceptibility genes for complex diseases. We have used a novel approach to identifying genes and characterizing SNP results that warrant further study in understanding the interaction between smoking and PD. For genomic research to be effective and socially just, we need to gain a better understanding of how the environment interacts with genes to modify complex disease risk.

Dedication

This dissertation is dedicated, with much love and deep gratitude, to the memory of my father-in-law, Balanna. He was both my inspiration and my biggest supporter, and I miss him every day. I am glad he saw me begin this process, even if he is not with me now, to see me complete my PhD.

Acknowledgements

"Somewhere, something incredible is waiting to be known." Carl Sagan

And while I was off on this journey of trying to find something incredible, I received a lot of help to get there....and I mean, a LOT of help! I couldn't have done this without all of you!! To start with, I would like to thank my supportive committee: Karen Edwards, Erika Blacksher, Cyrus Zabetian, Yvonne Lin, and Karl Hill, for their time, encouragement, and expertise throughout this process. I would also like to thank my department, IPHG for their unparalleled investment of time, and being truly caring of those who call the department home.

A special and huge thanks to Erika Blacksher whose extreme patience, unwavering support, and exquisite attention to detail helped me get to the point...I still have some work to do on getting to the point, but I am off to a good start.

A heartfelt thank you to Karen Edwards - an exceptional committee chair, mentor, and guide. You encouraged independent thought and were incredibly patient while I explored "academic rabbit holes"! I am extremely grateful for your steadfast support and very thankful to have worked with you all these years.

My sincere thanks to...

Lorelei, your uncommon kindness and warmth made the last few years at grad school a breeze ... well, as much as grad school can ever be a breeze!

Emilia, for putting up with me through the highs and lows of grad school while planning trips to far-off places and drinking endless cups of chai.

Ramya, for listening patiently to all my complaints, allowing me to vent without once saying 'I told you so', and pepping me up when I was low!

My sisters, Juls and Colleen, who have always believed in me and were my biggest cheerleaders through this process.

My husband Ram, and my son Prithvi, the two people in my life who've made success both possible and rewarding. I treasure your unconditional love and support. And last, but certainly not the least, I would never have gotten through the pressures of graduate school without my canine baby Ari, and his daily dose of love and hugs.

Table of Contents

List of Figures.....vi

List of Tables.....vii

Overview 1

Roadmap 1

Research Question 2

Social Justice in Genomics Research 2

Side A Complex Disease Genetics 3

Parkinson's Disease and GxE 3

In Conclusion 6

Social Justice in Genomics Research (Side B) 7

Premise 7

Scope of Research 7

Key Definitions 7

Social Institution 8

Social Justice 8

Social Goods 10

Why is Social Justice an Appropriate Framework to Evaluate Genomics Research? 11

Recognition 12

Table of Contents

Redistribution 13

Capabilities 14

Central Capabilities 14

Natural Environment as a Meta-Capability 15

Applying the Framework 18

Socially Just Genomic Research Methodologies 20

Genetic Research on Minorities will not Reduce Health Disparities 21

GxE, Recognition, and Capabilities 22

GxE, Redistribution, and Capabilities 23

Challenges, the Need for Increased Research and Tools to Develop GxE Methods 24

Lack of Detailed and Accurate Data on Most Exposures 24

GxE Differs by Developmental Life Stage, by Exposure Dosage, and by Genetic Variation Among Populations 25

Identifying Biomarkers that Constitute Direct Measures of Exposure are Critical to Evaluating GxE 25

Lack of Statistical Power 25

Lack of Replication 26

Small Effect Sizes and Lack of Statistical Significance 26

Epidemiologic Errors of Confounding and Misclassification 26

Table of Contents

Moral Imperative 27

Conclusion 28

Complex Disease Genetics: A Case Study of GxE in Parkinson's Disease (Side A) 29

Introduction 29

DNA Methylation and Phenotypic Variability 37

Methylation 38

Genetic Influences on DNA Methylation 38

Potential Mechanisms of Action: how SNPs Modulate Methylation 39

Parkinson's Disease 40

GxE, Coffee, and PD 47

Smoking 48

Exposure (Smoking) and Mechanisms of Action 49

Why are Causal Mechanisms Important in the GxE between Smoking and Genes in PD? 51

Rationale for GxE Analysis 51

Materials and Methods 53

Hypothesis 53

Materials 53

The NGRC Recruitment Process 54

The Diagnostic Process 55

Table of Contents

Diagnostic Criteria 55

Genotyping and Molecular Analysis 57

Methods 57

Framework for Selection of Genes and Variants for Testing 57

Step 1 58

Step 2 59

Selection of SNPs for the Analysis 62

Quality Control 64

Statistical Analyses 64

Results 66

Discussion 70

Missing Heritability of Complex Diseases and GxE 70

GxE Limitations 71

Exposure Data-Related Limitations 73

Additional Analyses 75

Functional Analysis 75

Tier 1 Gene CNTNAP2 and Annotated Results 77

Tier 2 Gene RARA and Annotated Results 77

Interpreting the LD Values 79

Table of Contents

RARA - A Susceptibility Gene 80

Next Steps 81

In Conclusion 81

References 83

Appendix ..... **Error! Bookmark not defined.**

List of Figures

Figure		Page
1.	Traditional models of GxE	35
2.	Regulatory model of GxE	36
3.	SNP rs2120200 and evidence for function	99
4.	SNP rs36030243 and evidence for function	99

Table of Contents

List of Tables

Table		Page
1.	Genes known to be differentially methylated by smoking	58
2.	Genes and SNPs by tier	60
3.	Location of gene and number of SNPs per gene	61
4.	Significance threshold	65
5.	Descriptive demographics of the NGRC dataset	66
6.	Tier 1 smoking analysis	67
7.	Tier 2 smoking analysis	68
8.	Tier 1 functional analysis	76
9.	Tier 2 functional analysis	76
10.	LD among the four SNPs in *RARA*	78
11.	Tier 1 sensitivity analysis	100
12.	Tier 2 sensitivity analysis	100

Overview

Roadmap

A substantive proportion of the promise of genomics rests on its potential ability to reduce burden of complex disease at a population level. Thus far, genomics research has not delivered on that promise, in spite of a vigorous research agenda and investment of public money. This dearth of benefit has implications for genomics research and research methods, and from a broader perspective, for social justice in genomics research. It is imperative for genomics research to provide benefits at a population level, and enable translational outcomes that can be delivered in population settings. Examining the aims, goals and promise of genomics research vis-à-vis relative lack of benefit is an important endeavor for the genomics research agenda. The task will help to better understand the challenges, limitations, and establish new avenues to prioritize for research exploration.

A potentially promising approach to address this problem within genomics research is the adoption of Gene-Environment interactions (GxE) research. In this interdisciplinary dissertation, we developed a framework by which to analyze social justice in genomics research, and using a case study of Parkinson's disease (PD), we demonstrated how GxE methods are useful in both moving genomics research forward in its endeavor on complex disease research, and more importantly meeting the demands of social justice in genomics research.

Research Question

This research attempts to answer the question of what would be a socially just genomics research agenda?''. The question is analyzed through a normative bioethical analysis, and a GxE analysis in PD genetics. We used both traditional and novel research methods, and tools that are decidedly exploratory to investigate GxE in PD. These types of exploratory analyses are needed to identify new avenues of research with the ability to create utility in PD genetics, as well as illustrate methods that meet the demands of social justice in genomics research by providing avenues for public health utility.

Social Justice in Genomics Research

All social institutions are subject to the demands of social justice, and in making the claim that publicly-funded genomics research is a social institution, we argue that genomics research is indeed subject to the demands of social justice. The normative analysis that follows examines the ways in which the current genomics research agenda may be responsible for misrecognition of vulnerable groups and unfair distribution of health and health outcomes. The analysis then evaluates what would be a socially-just research agenda, and highlights the importance, challenges, and opportunities of GxE research methodologies. GxE is a methodology with the potential to distribute benefit to the entire population, because it investigates the environmental burdens that are implicated in disease pathways.

Side A Complex Disease Genetics

The publicly-funded genomic research agenda in the US is heavily invested in complex disease genetics. But genotype-centric approaches have not yielded much utility in complex disease research. This is likely because of several reasons as discussed below.

a. Complex diseases are multifactorial. Genetics are not the only causal factor in disease etiology; the environment also plays a strong causal role in complex disease etiology. It is hypothesized that the environment is a strong regulator of gene expression and acts through gene-environment interactions (GxE) and epigenetic mechanisms to alter health outcomes in complex disease risk (Eichler et al., 2010; C. Liu, Maity, Lin, Wright, & Christiani, 2012).

b. Complex diseases have etiologies that tend to originate several years before phenotypic onset of disease, and it is crucial to gain a better understanding of the exposures that modify complex disease risk. Additionally, because some of these exposures may be modifying disease risk by interacting upstream in the pathway of disease pathogenesis, it is also crucial to gain a better understanding of the relevant timeframes to investigate exposures relative to disease onset.

Parkinson's Disease and GxE

PD presents as a perfect candidate for a case study of GxE. In addition to having a strong environmental component, as evidenced by high discordance in PD status among twins (Tanner et al., 1999a), PD also has strong environmental risk factors conferring both increased and decreased risk. While the association of these environmental factors with disease risk is well-

established, the causal mechanisms that confer these differences and the biological pathways on which they operate are not well-characterized as yet.

Some of the environmental factors that increase risk for PD are coterminous with the common characteristics that define vulnerable groups: i.e. ethnicity, low SES, rural living and rural exposures (Costa, Lunet, Santos, Santos, & Vaz-Carneiro, 2010; Kiyohara & Kusuhara, 2011a; Lix et al., 2010; Priyadarshi, 2001; Wright Willis, Evanoff, Lian, Criswell, & Racette, 2010). However, these are also groups on whom we have very little genetic information (Need & Goldstein, 2009a).

The environmental factors that consistently decrease risk for PD across all groups of the population include smoking (nicotine) and caffeine exposures (Costa et al., 2010; Kiyohara & Kusuhara, 2011a). Smoking is the exposure we chose to evaluate. Gaining a better understanding of the causal mechanisms by which smoking confers the decreased risk could aid translation of this area of research, regardless of the population studied. Identifying the causal mechanism could lead to therapeutic benefit using compounds with a similar mechanism, but without the undesirable effects of smoking (nicotine). The environment presents the mutable leverage point that can be modified to reduce disease burden or delay progression of disease.

Researching GxE in PD using environmental factors is socially just because it researches risk factors that impact all populations, and not just the subset on whom we have genetic data, and who have access to highly specialized medicine. Therefore, by researching the genetics of non-vulnerable groups within the population, but in interaction with an environmental exposure known to be associated with lower risk, could provide benefit to the entire population, including vulnerable groups, should this research come to translational fruition.

GxE Research Methods

GxE are notoriously hard to detect and are challenged by lack of statistical power as well as the selection of relevant exposures. Overcoming these challenges is a critical and unmet research need. It is necessary to invest in new methods that integrate multidisciplinary approaches to generate meaningful evidence, as well as robust methods by which to assign appropriate weights to the multiple lines of evidence.

In trying to overcome some of the traditional challenges of GxE, and include multidisciplinary approaches, this research model attempted to find new susceptibility genes involved in PD, identified from an environmental perspective and not just from disease phenotype. The intent was to find new susceptibility genes that are mutable (by environmental exposures) and in the causal pathway to PD risk.

To do so, we attempted to follow the environment into the body by identifying genes that are known to be impacted by the environmental exposure (smoking ever/never), as well as genes that are in the xenobiotic metabolism pathway of the environmental exposure (smoking ever/never). This set up an *a priori* hypothesis that somewhere along the network of genes known to be impacted by smoking is an interaction with gene networks involved in PD risk. Furthermore, one of the biggest effects of smoking is DNA methylation. To this end, we hypothesized that regulatory SNPs (rSNPs) on genes differentially methylated by smoking act as genetic determinants to epigenetic modification.

While the primary analysis (on the set of genes identified) was statistical (logistic regression), we further characterized the results for functional consequence and synthesized

multiple lines to evidence to identify PD susceptibility genes from the perspective of the environmental exposure. Using hypothesis-based approaches and layering functional information and tools over traditional genetic epidemiology methods, we used a multidisciplinary approach to interpret the data. This is a paradigm shift from siloed approaches to human biology, and highlights the opportunities that exist when different disciplines are integrated. Multidisciplinary approaches require further exploration because single-discipline specific approaches are a reductionist framework in complex disease research, and such approaches to research have not yielded much benefit.

In Conclusion

Genotypes are only as good as the environments they find themselves in (Gibson, 2008). Therefore, the necessity and the rationale to find the environmental factors that are in interaction with the genotypes.

Social Justice in Genomics Research (Side B)

Premise

The social institution of genomics research fails to meet the demands of social justice due to its paucity of research on environmental burdens that contribute to the excess burden of disease experienced by vulnerable groups. Genomics research should include genomic models that incorporate proximal and distal environmental burdens for it be socially just.

Scope of Research

Societies distribute health through complex processes, including scientific research, and specifically genomics research. This normative analysis examines the potential for unfair health distribution created by the genomics research agenda, and evaluates newer research methodologies that may be more socially just.

Key Definitions

Defined below are key terms used in this analysis.

Vulnerable Groups

Understanding what constitutes a vulnerable group is important to identifying and addressing social inequalities in health. Social inequalities in health (or health disparities) are caused by disproportionate burden of risk exposures faced by vulnerable groups. Vulnerable groups are different from at-risk populations. While at-risk populations have higher exposures to a specific risk factor, a vulnerable group is a subgroup of the population, who, on account of

shared social characteristics, are at higher risk of risks. Vulnerable groups are sub-groups within a population who share one or more of these common characteristics: low SES, race, ethnicity, gender, geographically rural, and other characteristics that may marginalize groups. This exposes them disproportionately to social, material, and environmental burdens that jeopardize health.

The notion of layering the risk conferred by shared social characteristics over the specific risk factor renders these groups vulnerable and their distribution of risk exposure has a higher mean than the rest of the population (Frohlich & Potvin, 2008). They are differences in health that are attributable to unjust social institutions and social policies.

This interpretation of vulnerable groups is all-encompassing, and includes minorities, marginalized groups, low SES groups, and rural populations. This is the frame of reference for how this analysis defines vulnerable groups: a population at risk of risks (Link & Phelan, 1995).

Social Institution

A social institution may be defined as a complex of positions, roles, norms and values lodged in particular types of social structures and organizing relatively stable patterns of human activity with respect to fundamental problems in producing life-sustaining resources, in reproducing individuals, and in sustaining viable societal structures within a given environment (Turner, JH, 1997). Social justice is the virtue on which social institutions are morally grounded.

Social Justice

Social justice is central to any account of a social institution, and is an important and necessary framework by which to evaluate social institutions. The role of the state in this endeavor is to support and enable social institutions in promoting social justice. This view is

supported by contemporary theorists of justice (Fleischacker, 2005).

In his *Theory of Justice*, Rawls argues that justice is the first virtue of social institutions (Rawls, John, 1971). Social justice requires institutional systems and governmental policies to act in concert and enable fair institutions and socially inclusive institutional frameworks.

Social justice is a commitment to the view that a fair distribution of rights, duties, and natural and social resources should be made across society, and this requirement of a fair distribution should be pursued by those in positions of societal, political and organizational responsibility. It pays attention to, and is in solidarity with those who are disadvantaged and excluded in society and posits that socially just structures are vital and should be maintained as a key to achieve social justice (Disney, Julian, Baldry, Eileen, Calma, Tom, & Briskman, Linda, 2011).

There is no single agreed upon definition of social justice, and there has never been, and is not now, agreement on the nature of justice (Hampton, 1997). For the purposes of this research, social justice is defined in terms of, and bound by the two overarching concepts in contemporary social justice: the principles of fair distribution and equal recognition. Justice requires both fair redistribution and equal recognition. The two principles form co-fundamental and mutually irreducible dimensions of justice (Fraser, 2003). Both recognitional harms and distributional harms cause health disparities, or differences in health that are attributable to unjust social institutions and social policies.

The working definition of social justice for this analysis draws on both recognition and redistribution as dimensions of social justice whose demands ought to be met if genomics research is a socially just social institution. Further, this definition also draws on a capabilities

approach; the idea that genuine opportunities for health (a life free of preventable morbidity and premature mortality) should be among the goods subject to a fair distribution. The environment (socioeconomic conditions and the exposures that they predispose) are causal in health outcomes, and therefore, are also among the goods that ought to be subject to a fair distribution to provide genuine opportunities for health. *Social justice applies to social institutions in distributing social goods.*

Social Goods

Justice theories offer quite different interpretations of which goods should be subject to a fair distribution. Rawls, in his notion of justice as fairness, makes the claim that the social good that requires a fair distribution are adequate primary social goods to make it possible to meet ever person's basic needs (Rawls, John, 1971). Faden and Powers posit that justice requires a fair distribution of good outcomes, specifying six "essential dimensions of well-being" that all persons should have a basic minimum of (M. Powers & Faden, 2008). Sen advances the notion that the social goods that ought to be subject to a fair distribution are neither resources, nor outcomes. Instead, they are capabilities. Under this theory, justice as capabilities is the provision of genuine opportunities “to be and do things people have reason to value" (Sen, 1995). Sen proposes that any claim to equality must take account of the diversity of human beings and their varying characteristics. Sen observes that the difference between most all contemporary ethical approaches to social arrangements and social justice lies not in whether they all demand equality of something, but rather in what "space" they propound equality. Sen argues that we should be concerned with people's capabilities rather than with their resources and makes the claim that a fair distribution of genuine opportunities enables social justice in

the space of capabilities.

Why is Social Justice an Appropriate Framework to Evaluate Genomics Research?

Publicly funded (health and disease-related) biomedical research, including genomics research, is tasked with "protecting and improving health" (National Institutes of Health, 2014). Genomics research, then, having been established for a common purpose (protecting and improving health), and regulated by a set of norms (policies that govern research and paid for by public dollars) meets the definitional criteria of a social institution.

In addition to being a social institution, genomics research, by virtue of being in health research is a part of the health system (the sum total of all organizations involved in improving and distributing health). According to the United Nations, health systems act as powerful drivers of inclusion or exclusion and have the potential to impact population health as a result of the policies and standards of practice they adopt (UNRISD, 2013). Faden and Powers argue that health is an essential dimension of well-being and state that social justice is the foundational moral justification for public health.

If publicly funded genomics research is a social institution tasked with protecting and improving population health (and improving public health), and, social institutions are subject to the demands of social justice, then it follows that genomics research outcomes should yield a fair distribution of important social goods. Therefore, it is both appropriate and necessary to evaluate genomics research through the lens of social justice.

Genomics Research Agenda and Social Justice

Inequalities in the two key constructs of social justice: a maldistribution of important social goods and misrecognition are causally implicated in the creation and/or exacerbation of health and health care disparities. Focusing on how genomics research agenda could be contributing to these disparities, we identify how each of the principles of social justice, and the capabilities approach to social goods map out within the genomics research agenda.

Recognition

Recognition is the dimension of justice that concerns itself with the politics of identity, and the demand for equal recognition mandates that assimilation into dominant cultural norms or the majority is not the price of equal respect (Fraser, 2003). Harms of misrecognition have been causally implicated in health disparities.

From the perspective of genomic research, recognition concerns itself with what we research, who we conduct research on and the reasons why, as well as how we prioritize research funding allocation. The extent to which these differences within the research agenda are attributable in creating new disparities and/or exacerbating existing disparities, they are deemed unjust. What this demand of justice requires in the space of genomics is research that translates to reducing misrecognition and ameliorating health disparities. When this demand is met, it will enable capabilities (that research is capable of providing) by providing genuine opportunities for affiliation that promote better health. Reducing harms of misrecognition

alone will never be sufficient in the pursuit of health equity. To be effective, we must also reduce harms from maldistribution.

Redistribution

Redistribution is the dimension of justice that demands a fair distribution of benefits and burdens of important social goods among all members of society. The notion of fair distribution of benefits and burdens requires us to ask the question "who should get what, and how will benefits be allocated"? In theories of justice, health is often considered important; either as among the important goods to be distributed fairly or as a prerequisite for the goods to be distributed fairly.

Vulnerable groups within the population face severe maldistribution of important social goods, including lack of beneficial environments as well as increased exposures to burdensome environments, both of which are causally implicated in poor health. These disproportionate environmental burdens are also causally implicated in complex disease etiology and are thought to impact disease risk by gene-environment interactions (GxE) that alter gene regulation.

Given that translational tools of genomic research are thought to have the ability to impact health and healthcare outcomes, and also given that the environment plays a primary role in the etiology of complex diseases, what this demand of recognitional justice requires in the space of genomic research is research that evaluates environmental burdens alongside genetics, and research that guarantees an equitable distribution of important social goods - the

fruits of genomic research. When this demand is met, it will enable capabilities (that research is capable of generating) by providing genuine opportunities for good health.

Capabilities

Capabilities combine elements of both principles of recognition and redistribution. From a capabilities perspective, being healthy is a desirable state of existence, and is a state of being that people have reason to value. From the point of view of genomics research as a social institution engaged in health research and social justice, this requires genomics research to fulfill its charge, enable a fair distribution of the benefits from genomic research, and create genuine opportunities for better health through its research agenda setting and translational tools of research. Iris Marion Young articulates that the role of social justice is to evaluate if the choices and the range of options provided to individuals by social institutions are fair (Young, 2013). While she does not address capabilities specifically, the range of options she refers to may be likened to capabilities created by social structures and social institutions.

Central Capabilities

To identify what specific capabilities ought to be enabled by genomics research in its role as a social institution subject to the demands of social justice, we apply Nussbaum Central Human Capabilities in which she articulates ten core capabilities that are central to human existence. Martha Nussbaum holds the view that although personhood may be shaped by differing circumstances, the underlying core of personhood and human flourishing is something that all humans share. Nussbaum moves the 'Capabilities Approach' further by identifying a list of 10 capabilities she calls Central Human Capabilities based on the activities they enable (M. C.

Nussbaum, 2011). Nussbaum views the list as an outgrowth of a process of critical normative argument grounded in human dignity. The list of capabilities is what she deems essential for human flourishing.

Of the ten capabilities Nussbaum identifies, #1, # 2, and #7 (Life, Bodily Health, and Affiliation) are three of the central capabilities are relevant to this critique of genomics research through the lens of social justice.

1. *Life*. Being able to live to the end of a human life of normal length; not dying prematurely, or before one's life is so reduced as to be not worth living.
2. *Bodily Health*. Being able to have good health, including reproductive health; to be adequately nourished; to have adequate shelter.
3. *Affiliation.* a) Being able to live with and toward others, to recognize and show concern for other humans, to engage in various forms of social interaction; to be able to imagine the situation of another. Protecting this capability means protecting institutions that constitute and nourish such forms of affiliation, and also protecting the freedom of assembly and political speech.

   b) Having the social bases of self-respect and non-humiliation; being able to be treated as a dignified being whose worth is equal to that of others. This entails provisions of non-discrimination on the basis of race, sex, sexual orientation, ethnicity, caste, religion, national origin and species.

<u>Natural Environment as a Meta-Capability</u>

Nussbaum's Capabilities do not tackle issues around how society distributes environmental benefits and burdens. While Nussbaum's list of capabilities does explicitly

include the natural environment in capability #8, Being able to live with concern for and in relation to animals, plants, and the world of nature, it does however, define the natural environment as instrumental to just that specific capability and not in the context of how it impacts other capabilities. Control over one's environment (Nussbaum's 10th central capability), while identifying the environment as a capability, defines the environment from only a political and material perspective, and excludes the natural environment. The natural environment is not referenced in any of Nussbaum's core capabilities.

There is cause to extend capabilities approach to include the environment as a meta-capability (Holland, 2008). The natural environment, however, is a prerequisite to all capabilities, and most certainly is a prerequisite for being able to live a life of normal length, being able to have good health, and being able to have control over one's environment. Because the environment is instrumental to good health and is also causal in complex disease outcomes, it is a pre-requisite to health. Holland articulates the environment as a capability from natural environment and environmental justice perspectives and makes the claim that we ought to factor environmental benefits and burdens when making public policies that promote capabilities. She goes on to point out that prominent justice theorists have not factored in the instrumental value of natural environment (to health and disease) because the environment has traditionally been considered indivisible and therefore not subject to unequal distribution. However, we know this to be untrue. The evidence that environmental burdens are greater among vulnerable groups is well-documented and indisputable - vulnerable groups disproportionately inhabit environmentally unsafe neighborhoods such as inner cities, reservations, near coal-burning power plants, and near hazardous and polluted sites (Chang &

Lauderdale, 2009; Mechanic, 2007; Rubin, Colen, & Link, 2010; Singh, Azuine, & Siahpush, 2013). Holland points out that these environmental factors are also shaped by the same maldistribution in wealth and power. Holland argues that Nussbaum's capabilities approach should be expanded to establish the importance of the natural environment's instrumental value as a condition of justice, specifically that certain ecological conditions are meta-capabilities: necessary for all the capabilities on the list of 'central human functional capabilities'. She proposes adding 'sustainable ecological capacity' as a meta-capability that enables all capabilities on Nussbaum's list and involves 'being able to live one's life in the context of ecological conditions that can provide environmental resources and services that enable the current generation's range of capabilities; to have these conditions now and in the future'.

Defining the Natural Environment

Because there is a strong link between natural environment and socio-economic/built environment; any process of altering the environment that violates and/or diminishes the instrumental value of the natural environment - be it socio-economic, industrial, political, or other man-made processes constitutes an inequitable distribution of the natural environment. Environmental burdens and environmental exposures with respect to genomics include social, behavioral, man-made and natural exposures.

Capabilities Approach is Useful in a Critique of Genomics Research

In addition to the three core capabilities, this argument for the environment as a meta-capability brings the natural environment within the purview of capabilities approach, and is

useful in examining inequitable distribution of environmental burden, and inequities in health causally attributed to these differential burdens.

Therefore, a capabilities approach is useful in a critique of genomics research because of what 'capabilities' are. People have capabilities - freedoms to achieve valuable ends only when they have genuine opportunities to avail. In the context of a critique of genomics, research that is solely focused on biological variation can be critiqued for failing to elucidate the environmental burdens known to be causal in the etiology of complex, multifactorial diseases.

To sum up, the framework outlined for this normative analysis utilizes capabilities and relevant central human capabilities to evaluate social justice in genomics research agenda. Furthermore, the framework identifies the relevance of and utilizes the environment as a meta-capability, a capability that enables all capabilities.

Applying the Framework

Complex human diseases, a strong area of research interest for genomics research, have been shown to be heavily influenced by environmental exposures and are thought to contribute to disease phenotypes by interacting with genes and altering gene expression and regulation. The environment may be defined as all non-genetic contribution to the variation within a disease phenotype. Disease genetics in complex, multifactorial diseases have met with limited success and this has spawned the idea that there is missing heritability in these diseases (Manolio et al., 2009). While a number of factors have been hypothesized to account for the missing heritability, including gene-environment interactions, genomics research is still largely

focused on investigating only the genetic drivers of disease and mostly in White populations (Need & Goldstein, 2009b).

Advances in fields such as pharmacogenomic have highlighted the role of genetic variants in response to drugs and are now being used clinically to identify adverse drug reactions and effective dose. Genetic screening for penetrant and causal variants (such as the BRCA 1/2 tests) also utilize genetic variants to identify risk susceptibility. Translational tools of genomics are beginning to become available to those with access to such tools. If these research avenues are successful, they have the potential to yield several translational tools in healthcare and clinical settings, yet these tools may not generalize to the population at large if the translation does not include the genetic drivers of disease for non-white populations. Additionally, these tools are available primarily in clinical and urban settings. This excludes vulnerable groups from benefitting from these tools because these are also the groups that often lack healthcare access. Genomics research is focused on the potential to deliver healthcare oriented end-points, but we need public health-oriented end points for it to be socially just and serve the needs of vulnerable groups within the population.

Furthermore, healthcare makes only a minor contribution to health and healthcare disparities do not explain health disparities (Braveman, 2006). We know that even in countries that have universal healthcare, the gradient in health mirrors the socio-economic gradient (Marmot, 2005).

The major drivers of health disparities are social conditions (Bridget C. Booske, Jessica K. Athens, David A. Kindig, Hyojun Park, & Patrick L. Remington, 2010). These are commonly referred to as the social determinants of health (Krieger, 1999; Link & Phelan, 1995; Marmot,

2005). These social determinants include harmful environmental exposures, which have been shown to be causal in the etiology of complex multifactorial diseases. Genomic research is heavily invested in identifying genetic causes to disease etiology, and not as much in interactions research in complex diseases. While it may be countered that we do not need genetics to develop policies that reduce harmful exposures, it may also be countered, and we do counter, that understanding disease etiology is critical to establish optimal leverage points for intervention, as well as prioritizing resource allocation and policy setting.

Using this framework, it is evident that the demands of social justice and principles of redistribution and recognition are unmet by genomic research. It has not made a dent into remedying differences in health caused by race and class, and neither have the benefits of genomic research been universal and equitable. From a capabilities perspective, the current genomic research agenda does not fulfill its role in elucidating causal pathways that enable the environment as a meta-capability. Lack of this meta-capability translates to reduced capabilities in the three core functional capabilities identified earlier (life, bodily health, and affiliation) as being central to social justice in genomics research.

Socially Just Genomic Research Methodologies

To do so, genomics research will have to model environmental exposures using interaction models to identify causal mechanisms by which the environment contributes to the disproportionate incidence of complex diseases among vulnerable groups. Identifying those pathways of interaction will present with leverage points to intervene on in the disease, both from proximal and distal perspectives.

There is a genomics research methodology with the potential to move the needle on health disparities, and it is one that models the environment in the disease etiology. Gene-Environment interactions research (GxE) has the potential to elucidate biological mechanisms and pathways and identify how the environment manifests in the human body to cause disease. It can assess the substantive portion of risk caused by interactions of genes and environment. The environment manifests in human biology by regulating and altering gene expression at a cellular level. This is the biological basis by which the environment is causal in disease pathways. Therefore it is crucial that we investigate the genome simultaneously with the environment to identify the mechanisms by which the environment alters regulation.

Genomic research will meet the demands of social justice by engaging in research that identifies GxE causal pathways to disease risk caused by exposure to risk environments. To do so would require us to turn the investigative question from starting with disease phenotypes and identifying genes engaged in the pathway between the genes and disease to starting with environmental exposures and following the environment into the body to ascertain how the environmental exposure exerts influence on the body in its interaction with genes.

Genetic Research on Minorities will not Reduce Health Disparities

Funding agencies such as the NIH stress the importance of minority participation in genomic research as a pathway to reduce health disparities, the presupposition being that such participation will reveal genetic variation that is attributable to increased disease risk. Pursuant to this idea that genomic research on minorities will reduce health disparities, is an increased call for minority participation in health disparities research to identify genetic variability that is causal in disease mechanisms. While this will fill in gaps in knowledge on variations that exist in

minority populations, and potentially provide insight for clinical applications in healthcare, it will not move the needle on health disparities in vulnerable groups.

In order to impact health, we need to research the mutable factor and the primary factor; the environment. Researching GxE offers that option. It does so by elucidating the effect of the environment on complex disease outcomes, and identifying the biological pathways that environmental exposures operate on. The most promising way to reduce health disparities is to intervene on socially controllable determinants of health, among which are features of the natural and social environments. Although there is strong evidence for the correlations between environmental exposures and complex disease outcomes, there is a need to generate evidence for biologically-based causal associations between the two. Genomics research is the most mature and the most funded of all the 'omics' research platforms, and is best positioned to inquire the intersection between human biology and the environment. Therefore, in addition to the call for minority participation in research as a pathway to reduce health disparities, we ought to also be making a strong a call for increased research modalities and investigations that factor in the environment as well as genetics in research on complex diseases.

GxE, Recognition, and Capabilities

By investigating the environmental conditions that impact everyone, research on GxE is inclusive of all races and ethnicities because it clarifies the impact and biological networks of the exposure regardless of the population under study. This creates genuine opportunities for affiliation for vulnerable groups to be recognized for their disproportionate burdens of harmful environmental exposures.

Investigating the environment and modeling it into disease genetics prevents stigmatizing already vulnerable groups of people based on their genes and enables the capability for affiliation. Complex diseases are by nature multifactorial and investigating genetics of complex diseases, especially behavioral genetics, poses the challenge of inadvertently stigmatizing large groups of people based on their genotypic association with disease phenotype. This is not only unjust, but also unwarranted because these genotypes on their own usually do not cause disease outcomes. They only do so in interaction with the environment.

GxE, Redistribution, and Capabilities

In addition to investigating proximal (downstream) environmental factors that are causally implicated in complex diseases, GxE research methodologies have the potential to also investigate distal (upstream) environmental factors implicated in complex diseases, and it is possible that such research may identify leverage points to intervene on before disease outcomes or leverage points that alter/revert disease progression. The interventions could be clinical, public health measures, policy measures, altering the reach of specific exposures by altering standards of exposure such as no observed adverse effect levels (NOAEL) or low observed adverse effect levels (LOAEL) or banning toxins known to interact in disease pathways. Research of this nature will have a positive impact on health and enable environment as a meta-capability as well as core capabilities to life and bodily health.

Challenges, the Need for Increased Research and Tools to Develop GxE Methods

However, research of this nature also requires funding for development of new and more accurate methodologies, new biomarkers to evaluate, creation of large consortia, standardized exposure evaluation methods , and of course, more GxE research. There are methodological challenges that need to be overcome to successfully research GxE in complex diseases. The genetics of complex diseases have been, and continue to be a challenge to unravel. While some of these challenges are also faced by epidemiologists investigating only genetic risk factors, GxE research also poses some unique challenges that need to be overcome in order for it to be successful in evaluating the interplay.

Lack of Detailed and Accurate Data on Most Exposures

Exposure evaluation questionnaires are very useful tools in assessing risk. However, there are limitations in their ability to accurately collect exposure data. They are usually concise in order to minimize time spent filling them, but that also means that they lack the level of detail required for these types of analyses. It is now known that most complex diseases have etiologies well upstream of phenotypic onset of the disease, in which case documenting exposures that occurred several years in the past, and dated relative to disease onset, is vital to understanding the interaction between the environmental exposure and the disease. This type of documenting is strongly subject to recall bias. Additionally, health behaviors that are known to be risky will be subject to reporting bias, and we may not be able to document it accurately using exposure questionnaires (Coughlin, 1990; Stewart, 1999).

GxE Differs by Developmental Life Stage, by Exposure Dosage, and by Genetic Variation Among Populations

It is well established genetic variations are causal in expression of GxE within populations (Purcell, 2002). There are also other factors that cause differences in GxE expression. Genes are programmed to turn on/turn off based on developmental stages. This sets up critical stages during which the effect of the exposure is different (Wright & Christiani, 2010). The effects of GxE also vary by exposure dosage (Garte, 2006). These differences in the expression of the interactions highlight the need to identify optimal time points, accurate exposure dosage, and the need to research on all populations in order to successfully identify the biological basis on which these interactions operate.

Identifying Biomarkers that Constitute Direct Measures of Exposure are Critical to Evaluating GxE

This is a challenge and an unmet need in GxE research. Thus far, there have been relatively few studies that have assessed associations between biomarkers of biologically effective exposure dose and disease status, and most studies in the molecular epidemiological literature have been transitional studies assessing associations among exposure, genetics, and biomarkers (Hulka & Margolin, 1992; Rundle & Schwartz, 2003).

Lack of Statistical Power

Although lack of statistical power is an issue that is common across most all genome-wide association studies (GWAS), it is amplified in GxE studies, because cases and controls are each stratified by two factors: genetics and environment. Gathering data that are qualitatively

reliable, and increasing numbers of participants in such research are both required to improve the power of GxE to detect effects.

Lack of Replication

To date, GWAS studies have challenges with replication of results, and GxE studies are no exception to that. Usually, in these studies, the variants identified are in linkage disequilibrium (LD) with the causal variant/variants, spurious associations, lack of definitional clarity around the phenotype being looked at, and effect modification. A GxE study that is well-designed could reduce the effect modification by accounting for and analyzing those exposures that cause the modification.

Small Effect Sizes and Lack of Statistical Significance

A large proportion of the research show only modest effect sizes and p-values that border significance. The contention in a GxE analysis is that any evidence for a multiplicative effect is evidence for interaction and if the study is well-designed and adequately powered, then the evidence base is rigorous and the results are valid.

Epidemiologic Errors of Confounding and Misclassification

Errors of confounding and misclassification are common to all epidemiologic studies, and robust study design is a prerequisite for all such studies.

Investment in GxE research and development of new tools/methodologies that integrate varied scientific disciplines - such as genetic epidemiology, bioinformatics, systems biology, research on biomarkers, exposure evaluation, environmental sciences and toxicology are required to yield research outcomes that provide benefit. There are opportunities and challenges involved in developing novel study designs and tools to analyze in GxE. There is also

the need for investments in other aspects of GxE research, including, but not limited to developing alternative approaches to exposure assessment(Ghazarian et al., 2013; Hutter, Carolyn & Mechanic, Leah, 2013). Newer methodologies that rely on multiples lines of evidence such as those generated by systems biology and pathway approaches (methods of non-agnostic inquiry), and characterizing SNP (genomic) data from interaction models for functional consequence provide value, and can be used in addition to the statistical methods and statistical significance. This ties in to current research funding policies that prioritize the reuse of pre-existing data, and also prioritize innovative methods to re-examine the existing data (Paltoo et al., 2014). To sum up, a number of approaches are needed: novel analytical tools that can accurately and reproducibly characterize and measure both phenotypes and environmental exposures; population biobanks that enable prospective genotype, phenotype, and environmental data to be collected from large cohorts; and effective strategies for integrating all these data for meaningful analyses(Thorisson, Muilu, & Brookes, 2009).

Moral Imperative

There is a moral imperative to prioritize funding and development of research and tools for use in GxE to develop the evidence base that will allow us to tackle fundamental causes and health disparities through the strong biomedical policy initiatives that are already established with the funding agencies. The call for increased investment in GxE research requires the explicit recognition that the current advances in genomic research will not move the needle on health disparities, not unless we develop a strong body of evidence on the link between human biology and environmental burdens, the causal link that predisposes vulnerable groups to

increased risk through GxE. This is a moral imperative required of the genomics research agenda because policy changes are more often than not small and incremental shifts. The shifts require a substantive evidence-base across many disciplines to be considered as priorities for implementation.

Conclusion

If genomics research prioritizes GxE research in complex, multifactorial diseases with environmental points of entry into disease pathways, it has the potential to create capabilities or genuine opportunities at a population level for a longer life without preventable morbidities, better bodily health, the bases for affiliations, as well as the meta-capability of environment. GxE is an important pathway by which the environment manifests in the human body and leads to health outcomes. It also represents the intersection between the realms of social justice and genomics research because GxE and the subsequent gene regulation that follows are mechanisms that constitute the embodiment of environmental exposures in our biology.

Complex Disease Genetics: A Case Study of GxE in Parkinson's Disease (Side A)

Identifying genes that are known to be differentially methylated by smoking provides a starting point to evaluate SNPs on a group of genes- either because they are in the clearance pathway of nicotine or because they are known to be differentially methylated by smoking. This provides a way to assess known risk factors by following the environmental exposure into biological functioning. This has the potential to identify new susceptibility genes involved in PD, identified from an environmental perspective and not just from disease phenotype.

Thus, the overall goal of this project is to: Identify those genes that are differentially methylated by smoking, and genes involved in the metabolism of nicotine, and to perform an evaluation of GxE in Parkinson's disease (PD). We hypothesize that regulatory single nucleotide polymorphisms (SNP's) underlie a proportion of the regulatory differences in gene expression caused by methylation between cases and controls in PD, and analysis of SNPs on these genes could capture genetic differences between cases and controls in interaction with the environmental exposure (smoking ever/never).

Introduction

The completion of the Human Genome Project (HGP) and the HapMap project facilitated the growth of genome- wide association studies (GWAS) as a methodology to understand the genetics of complex diseases. To date, GWAS studies have identified thousands of common variants implicated in disease traits but most of these variants are in non-protein coding regions of the genome. Additionally, the SNPs that have been associated with disease

traits have modest effect sizes. And while complex traits differ in their underlying genetic architectures, for many common disorders the predominant pattern is that of many loci, each with small effects on phenotype. The median odds ratio for the disease-associated SNPs is (OR) 1.33 (Hindorff et al., 2009; B. E. Stranger, Stahl, & Raj, 2011). The lack of large effect sizes does not invalidate the association, but it does indicate that the strength of the association is typically not robust. GWAS does, however, offer ways to rigorously catalog common genetic variation and document the SNPs observed to be associated with disease traits.

"One of the great hopes for GWAS was that, in the same way that huge numbers of Mendelian disorders were pinned down at the DNA level and the gene and mutations involved identified, it would be possible to simply extrapolate from single gene disorders to complex polygenic disorders. That really hasn't happened. Proponents will argue that it has worked and that all sorts of fascinating genes that predispose to or protect against diabetes or breast cancer, for example, have been identified, but the fact remains that the bulk of the heritability in these conditions cannot be ascribed to loci that have emerged from GWAS, which clearly isn't going to be the answer to everything" - Sir Alec Jeffreys, ESHG Award Lecturer 2010 (Visscher, Brown, McCarthy, & Yang, 2012).

There are several reasons that GWAS on complex diseases have not yielded the same quality of results as the genomic research on Mendelian disorders. To start with, the biology of complex diseases is multifactorial. Genetics are not the only causal factor in disease etiology, as the environment also plays a strong causal role in complex disease etiology. It is hypothesized that the environment is a strong regulator of gene expression and acts through GxE and epigenetic mechanisms to alter health outcomes in complex disease risk. Moreover, even

within the genetics portion of complex diseases, complex diseases are typically polygenic (unlike monogenic Mendelian diseases) (Weeks & Lathrop, 1995). This means that genes at many different loci contribute cumulatively to disease risk. To add to the biological challenges are the statistical and epidemiological challenges. GWAS are agnostic by nature and subject to high penalties of adjusting for multiple comparison. Also, in terms of characterizing the associations from GWAS, the results do not evaluate how these disease associations impact biological functioning. And finally, SNPs found associated in GWAS with disease traits are usually in non-protein coding areas of the genome such as introns and intergenic regions. There are two hypotheses that explain the association of these SNPs with disease traits. The first hypothesis is that these are not causal SNPs, but tag SNPs that are in linkage disequilibrium (LD) with the causal SNP. The second hypothesis is that these intronic/intergenic SNPs are hallmarks of regulatory SNPs (rSNPs) that act in concert to engage in epigenetic mechanisms.

"The Encyclopedia of DNA Elements (the ENCODE project) was designed to pick up where the HGP left off. Although that massive effort revealed the blueprint of human biology, it quickly became clear that the instruction manual for reading the blueprint was sketchy at best. Researchers could identify in its 3 billion letters many of the regions that code for proteins, but those make up little more than 1% of the genome, contained in around 20,000 genes - a few familiar objects in an otherwise stark and unrecognizable landscape. Many biologists suspected that the information responsible for the wondrous complexity of humans lay somewhere in the ‘deserts’ between the genes. ENCODE, which started in 2003, is a massive data-collection effort designed to populate this terrain. The aim is to catalogue the ‘functional’ DNA sequences that lurk there, learn when and in which cells they are active and trace their effects on how the

genome is packaged, regulated and read" (Maher, 2012).

In this landscape of genetic research, GWAS is the most mature platform with a lot of research investment, albeit with its own set of limitations, and then there are the newer tools of bioinformatics that use ENCODE and have immense potential, but are not as tried and tested.

The Need for Non-Reductionist Approaches

The spotlight on GWAS, its capabilities and limitations, underscores the need for more nuanced and non-reductionist approaches to genomic research of complex diseases. These include adding epigenetic information (using bioinformatics) around the genetic SNP data to understand the function and annotate the signal from GWAS. Epigenetics and gene regulation are layers of complexity that reside on top of the genetics. While epigenetics on its own can provide information that accounts for some of the complexity, there is more meaning that emerges when epigenetic information is positioned against the backdrop of genetics and genetic variation. While genomic research using agnostic models such as GWAS have not yielded much by way of identifying or understanding genetic risk, neither have single gene approaches using a monogenic model of inheritance. While there are single genes strongly linked to disease risk in PD, explain a proportion of the risk, and are implicated in the familial forms of the disease, the bulk of the genetic susceptibility in complex diseases likely arise from polygenic models of risk, especially in the sporadic forms of the disease. It is also critical to model the environment and use GxE models to studies of complex diseases. Pathway approaches are non-agnostic, hypothesis-based and look at disease phenotypes from the lens

of what sets of genes along what pathways are turned on or turned off. Pathway approaches also designate candidate genes of interest, but these genes are based on biological plausibility and interact with each other along specified pathways.

One way to combine all of these concepts: epigenetic modification, GxE, pathway approaches; is to model a regulatory approach in a GxE association study, and follow a specific environmental exposure in to the body, and identify groups of genes that are known to be impacted in a regulatory manner by the exposure. This information has the potential to identify new susceptibility genes for disease phenotypes, and could be used to characterize the missing variation in the disease phenotypes.

Towards Reuse of Pre-existing Data

Since the HGP, publicly-funded health research in the United States has allocated several tens of millions of dollars in research funding towards conducting GWAS research. Some of these studies also have environmental information (collected via questionnaires) completed by the research participants. The data from these research studies have been analyzed for genetic associations but there have been few studies that have utilized pathway approaches or bioinformatics with an emphasis on the environmental data. These new approaches may permit the identification of susceptibility genes using pre-existing data, genes that would not have been identified by a strict GWAS approach and contribute to a better understanding of the disease process. This process constitutes data mining, in that it seeks to extract patterns from the data, but at the same time is based on *a priori* hypothesis of biological plausibility that guide what genes are selected for the analysis. Additionally, reuse of pre-existing data is a

policy priority for funding agencies, as are innovative methods to re-examine the data (Paltoo et al., 2014).

Gene-Environment Interactions Research and Regulatory GxE

The Need for GxE Research (GxE Models and the Regulatory GxE Model

The environment acts to modify complex disease risk, and it is widely acknowledged that there is a need to identify these interactions in order to stratify disease risk by environmental exposure and genotype status in genetic studies of complex diseases.

Gene-environment interaction research is an extension of GWAS, but one that models the environmental exposure. The environment and the SNPs are modeled as main effects and their combined interaction is modeled as a multiplicative joint effect. It is subject to the same challenges as GWAS, but statistical power is further impacted and compounded by the relative lack of accurate and detailed exposure measurements.

The environment exerts its influence on genes through genetic and epigenetic effects. Both effects modify disease risk and health outcomes. Epigenetic effects such as histone modification and DNA methylation regulate gene expression. The environment is also in interaction with genes (genetic effects) through GxE. GxE occurs when the same environmental exposure has different effects on gene expression depending on the individual's genotype. This GxE model is the basis for pharmacogenomic interactions. We maximize the chances of identifying interactions if there is an exposure that is known to either increase or decrease risk of a specific disease or disease trait. Because the exposure is already well-characterized with respect to disease outcomes, it may be hypothesized that somewhere along the pathway of

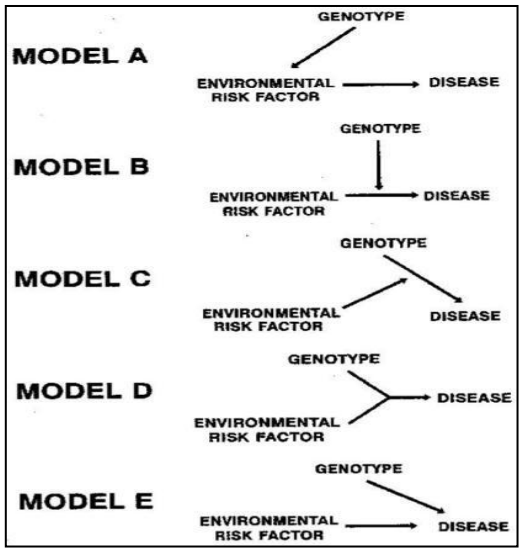

Figure 1: Traditional Models of GxE

metabolism of the exposure, and the effect of the exposure, lies an intersection with the disease pathway. If an exposure is consistently associated with altering disease risk, it must intersect biologically with the disease pathway to cause the altered risk. In thinking through this interaction model, and, considering the evidence that regulatory variation may be the key primary effect contributing to phenotypic variation in humans, we used the regulatory model of GxE as a framework (Idaghdour & Awadalla, 2013; Barbara E Stranger et al., 2007). It is similar to traditional models of GxE but slightly different in how it functions. The traditional models of GxE interactions are represented in Figure 1 (Ottman, 1996).

The model of interaction we hypothesize is a regulatory model of regulatory SNPs (rSNPs) represented below (see Figure 2). In this model, environmental exposures interact with rSNP to alter pre-existing states of regulation, and this interaction with the environmental exposure lies in the causal pathway to disease outcomes. The model differs from traditional models of GxE on account of its specific focus on regulatory SNPs, and also because it models the epigenetic modification caused by the interaction, which is then the proximal factor leading to the disease phenotype. This ties in well with our current understanding of complex disease etiology where environmental exposures that are associated with complex disease phenotypes

are well upstream (in terms of a timeline) from disease onset.

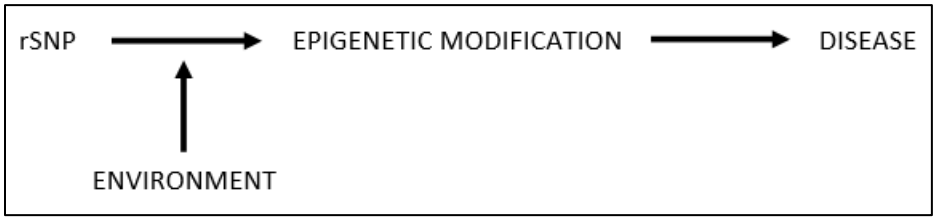

Figure 2: Regulatory Model of GxE

Under this regulatory model, it is hypothesized that regulatory SNPs (or rSNPs) located on loci that are also involved in epigenetic regulation, are likely interacting with environmental exposures through GxE. The rSNPs have functional value under regulatory circumstances, and allelic differences underlie epigenetic differences in regulation. Variations in rSNPs affect ability of a transcription factor (TF) to bind to DNA (Macintyre, Bailey, Haviv, & Kowalczyk, 2010). Epigenetic regulation of gene expression and/or gene function is different for different alleles of the rSNP. The environmental exposure interacting with the risk allele of the rSNP confers altered susceptibility to disease. This is one that models individual genotypes, the environment, and disease risk, but with a regulatory focus. Since it has been established that a number of SNPs found to be in association with disease traits are non-protein coding and likely regulatory in function, the importance of these SNPs in GxE and epigenetic modification is highlighted (Schaub, Boyle, Kundaje, Batzoglou, & Snyder, 2012; Wang et al., 2010). Regulatory SNPs may be thought of as genetic determinants to epigenetic sites.

This model only looks at SNP data, but with a focus on expanding it to look at regulatory SNPs. The hypothesis here is that epigenomic modifications vary by status of rSNP allele. The rSNPs polymorphisms alter the interaction with environmental exposures, interactions whose

mechanism of action is an epigenetic modification. These are SNPs that are upstream or downstream from genes, in the 5'- promoter region or 3'- tail, or in introns where they could alter recruitment of transcription factors.

DNA Methylation and Phenotypic Variability

DNA methylation patterns are important for establishing cell, tissue, and organism phenotypes, but little is known about their contribution to natural human variation. When differentially methylated CpG sites were examined, the differences in methylation could be partially traced back to genetic variation, suggesting that differentially methylated CpG sites serve as evolutionarily established mediators between the genetic code and phenotypic variability (Heyn et al., 2013). The same research also noted that one-third of the DNA methylation differences were not associated with any genetic variation, suggesting that variation in population-specific sites takes place at the genetic and epigenetic levels, highlighting the contribution of epigenetic modification to natural human variation. By extension then, two-thirds of DNA methylation are associated with genetic variation, and these methylation patterns likely establish cellular and tissue function. The research goal for this analysis was to identify the genetic determinants (regulatory variation) of epigenetic patterns (that are modifiable by environmental exposures) in complex disease genetics. To clarify, this research only looks at SNP data to find evidence for rSNPs, but using non-agnostic methods and multiple lines of evidence.

Methylation

Methylation is a chemical modification of DNA, and an epigenetic mechanism that occurs at CpG sites along the genome (the CpG notation represents a cytosine and a guanine linked by a phosphate along a linear sequence). Cytosines in these dinucleotides have a propensity for methylation and are converted to 5-methylcytosines. If cytosines within genes, or in the regulatory regions that control gene expression are methylated by an environmental exposure, it can alter gene function by turning off the gene (Miranda & Jones, 2007). Alternately, if an exposure reverses methylation, it can turn on genes that are silenced (Baylin, 2005). This change in state is catalyzed by the DNA methyl transferase (DNA MTase) family of enzymes coded for by three genes in mammals: *DNMT1, TRDMT1,* and *DNMT2* genes. Methylation modulates transcription of genetic information from DNA to RNA and, in this manner, can influence the expression of a given phenotype (normal traits or diseases). SNPs with regulatory function (rSNPs) located on CpG sites are also known as CpG SNPs. DNA methylation is more similar in monozygotic than dizygotic twins (Kaminsky et al., 2009). Moreover, age-related changes in DNA methylation during adulthood show familial clustering with estimated heritability of >70% (Bjornsson, 2008). This similarity could potentially be a result of shared genetic variation (rSNPs) at the CpG sites.

Genetic Influences on DNA Methylation

The contribution of genetic factors to DNA methylation may vary across individual CpGs. Some CpGs are located directly on SNPs and if the cytosines or guanines are mutated into other nucleotides, these sites may no longer be methylated. In these CpGs, DNA methylation behaves as a monogenic trait, being “high” in non-mutant homozygotes, “intermediate” in

heterozygotes and “low” in homozygotes. Effects of sequence variation can spread across neighboring CpGs and thus contribute to the observed correlated nature of DNA methylation at neighboring CpGs (Bell et al., 2011). For example, at a CpG located on rs10846023 (a T/C SNP), the level of DNA methylation is highly allele-specific (the C allele had a propensity for being methylated and the T allele did not); and this effect was also exhibited at nearby CpGs, spanning over 500 bp (Shoemaker, Deng, Wang, & Zhang, 2010).

Support for DNA methylation being in part determined genetically comes from GWAS testing genotype–phenotype associations between >600,000 SNPs and DNA methylation at >25,000 CpGs. These GWAS identified a large number of SNPs (~3,000) associated with the level of DNA methylation at various CpGs. Most of them were located within 2-kb regions of interrogated CpGs, but some were further apart or even on different chromosomes (Bell et al., 2011; Numata et al., 2012). This research also noted that given the large number of statistical tests typically performed in a GWAS of DNA methylation (i.e., 600,000 SNPs × 25,000 CpGs) and the need for correction for multiple comparisons, only SNPs with very large effects (explaining >20% of variance at a given CpG) have been reported, thus leaving out undiscovered SNPs with smaller but likely biologically meaningful effects.

Potential Mechanisms of Action: how SNPs Modulate Methylation

Other than creating or abolishing a CpG as a result of SNP variation at the specific locus, the mechanisms of how DNA variants modulate DNA methylation are not well understood. GWAS results suggest that some mechanisms may be regional, whereas others may be more global, or even genome-wide. The SNPs associated with DNA methylation at nearby CpGs are likely to exert regional effects. Regional effects may be related to specific sequence variants

interfering with the action of the DNA methylation machinery (Handa & Jeltsch, 2005). Research on type 2 diabetes-associated SNPs shows that about half of the SNPs examined (19/40) either introduced or removed a CpG site suggesting that introduction or removal of CpG sites may be a molecular mechanism through which disease-associated SNPs impact gene function via differential DNA methylation and contribute to disease phenotypes (Dayeh et al., 2013).

Parkinson's Disease

PD History

Loss of neuronal function is often regarded as a consequence of ageing. There are important public health implications in age-associated cognitive and motor declines. According to the World Health Organization, the proportion of people 60 and older is the fastest growing age group globally (United Nations, 2013). While this is a success story for public health policies that have helped reduce mortality from childhood and early-adulthood diseases, an older population poses challenges in terms of maximizing health outcomes and maintaining functional capacity. One such challenge for older populations is Parkinson's disease (PD). Eponymously named for Dr. James Parkinson, a British physician who first described the disorder in his book almost 200 years ago, it is a chronic and progressive neurodegenerative disorder of uncertain etiology and a prevalence of 1 % in individuals over the age of 65 (Jankovic, 2008).

Features of PD

The clinical features of the disease can be attributed to loss of neuronal function. Within the brain, PD is canonically characterized by a profound and selective loss of vital neurons in the mid-brain or mesencephalon, in an area called the substantia nigra. This area is involved in the production of dopamine. Dopamine is a chemical signaling molecule transmitting signals for controlling smooth physical movement and loss of dopamine is tied to motor impairment. Recent findings implicate mitochondrial dysfunction, oxidative damage, abnormal protein accumulation and protein phosphorylation as key molecular mechanisms compromising dopamine neuronal function and survival as the underlying cause of pathogenesis in both sporadic and familial PD (Thomas & Beal, 2007). At the level of an individual's behavior, these changes result in movement abnormalities, which are the major manifestations of the disease (Mazzoni, Shabbott, & Cortes, 2012). Rest tremor, bradykinesia, rigidity and loss of postural reflexes are generally considered the cardinal signs of PD. The presence and specific presentation of these features are used to differentiate PD from related parkinsonian disorders (Jankovic, 2008). As the disease advances, cognitive and other behavioral problems also arise.

Molecular Changes in PD

α-synuclein is a presynaptic protein that plays a crucial role in dopamine compartmentalization in the striatum (Yavich, Jäkälä, & Tanila, 2006). Normally seen in the brain as an unstructured soluble protein, α-synuclein plays an important role in PD etiology. The protein polymerizes to form insoluble fibrils (β-sheets) that ultimately coalesce into Lewy bodies (Tofaris & Spillantini, 2005). In most cases of PD, Lewy bodies are seen in the dead and impaired neurons. Lewy bodies are thought to be similar to aggregosomes (proteinaceous inclusion bodies comprised of several different proteins including α-synuclein and ubiquitin),

that form in a cell. Aggregosomes form in response to cellular stress, which causes impaired clearance of protein by the lysosomal system and the ubiquitin-proteasomal system. Whether Lewy bodies are causal to the disease or are associated with a protective cellular mechanism is research that is still in progress.

PD Genetics

Early research on PD predominantly pursued the association between PD and environmental factors such as neurotoxins. Genetic research on PD is a recent entrant to research on PD, but since the emergence of interest in PD genetics in the 1990's, there has been prolific research on the genetic underpinnings of the disease (Nussbaum, 1997). There are rare Mendelian forms inherited in families and the more common, non-Mendelian, sporadic forms of PD that still cluster in families but are without clear Mendelian segregation. Research has been successful in identifying specific loci and genes associated with Mendelian forms of PD. Genetic research has provided a deeper insight into our understanding of both familial and non-familial forms of PD (Hardy, Cai, Cookson, Gwinn-Hardy, & Singleton, 2006). However, assigning causal roles to genetic determinants in non-Mendelian forms of PD is lagging, as is information needed to improve prevention, detection and treatment of PD (Lesage & Brice, 2009).

PD Etiology

PD is traditionally classified into two forms: familial PD, where patients present with a strong family history of PD, and sporadic PD, with no family history or minimal familial clustering. Although the pathoetiology for both forms are thought to be different, both forms have been shown to include a genetic component. There are currently six genes associated with

heritable and monogenic forms of PD. Out of the six genes, mutations in *SNCA* and *LRRK2* are responsible for autosomal-dominant PD forms, and mutations in *Parkin*, *PINK1*, *DJ-1,* and *ATP13A2* are accountable for PD that displays an autosomal recessive mode of inheritance. They collectively account for about 30% of the familial and 3%–5% of the sporadic cases. Variants in several other genes (*MAPT*, *GBA*, *NAT2*, *INOS2A*, *GAK*, *HLA-DRA*, and *APOE*) have also been associated with an increased risk of developing PD. Of these risk susceptibility genes, β-glucocerebrosidase (*GBA*) is a well-validated PD-associated risk factor (Klein & Westenberger, 2012). Mutations on this gene increase the risk of developing PD and both homozygous and heterozygous mutations predispose to classical parkinsonism (Sidransky, 2006). However, mutations in all these genes represent only a small proportion of all known PD cases. Most research on sporadic PD indicates that the etiology is a combination of genetic and environmental factors, with some of the strongest evidence coming from twin studies that imply an environmental etiology or as yet unknown gene-environment interactions (Tanner et al., 1999b). While it does not appear that sporadic PD is caused by the deficiency of a single gene product inherited in a Mendelian manner, about 25% of all patients report having a relative with PD. Sporadic PD has been hypothesized to be caused by associated genes and susceptibility variants in those genes that confer greater risk when combined with other environmental factors. Idiopathic or sporadic PD is the second most common neurodegenerative disorder, second only to Alzheimer's disease.

Age is an important risk factor in the development of PD. Although thought of as a disease of old age, a small percentage of patients (about 5% of all cases) present with symptoms before the age of 60 years and the majority of these cases are caused by mutations

in an ever increasing list of genes which affect either protein metabolism or mitochondrial function, thus highlighting that dysfunction in either is sufficient to cause PD (Gasser, Hardy, & Mizuno, 2011). Gender is another risk factor with men being almost two times more likely to develop PD than women (Van Den Eeden, 2003). PD has a prevalence of approximately 1 million cases in the United States and about 60,000 incident cases are diagnosed each year (Parkinson's Disease Foundation, 2015). In spite of focused research, the pathoetiology of the disease is still elusive and relative contributions of genetic versus environmental factors are still in debate. However, by the time the disease manifests through the symptoms, more than 50% of substantia nigra (SN) dopamine neurons are lost (Ross et al., 2004). This suggests an etiology that is well upstream of the symptomatic disease manifestation timeline.

Environmental Exposure

Environmental Risk Factors Known to be Associated with PD

Epidemiologic studies have evaluated many environmental risk factors for PD, factors that are associated with both elevated risk and reduced risk.

Environmental exposures associated with increased risk of PD

There are several factors associated with elevated risk of developing PD, such as exposure to pesticides and rural living, well-water use, farming, SES, race, and ethnicity. Some of these factors are also associated with vulnerable groups.

Pesticides: A meta-analysis that evaluated 104 studies found that PD was associated with farming and the association with pesticides was highly significant in the studies in which PD diagnosis was self-reported. PD risk was increased by exposure to any-type pesticides,

herbicides, and solvents. Exposure to paraquat or maneb/mancozeb was associated with approximately a 2-fold increase in risk (Pezzoli & Cereda, 2013).

Rural living: A meta-analysis examining the association between PD and living in a rural area, reported a combined OR for rural residence of 1.56 (95% CI: 1.18–2.07) for all the studies, and 2.17(95% CI: 1.54–3.06) for studies performed in United States (Priyadarshi, 2001).

Well water use: A meta-analysis examining the association between PD and well water use reported a combined OR for well water use of 1.26 (95% CI: 0.97–1.64) for all the studies, and 1.44 (95% CI: 0.92–2.24) for studies done in United States. This could be the effect of pesticide runoff, persistent organic pollutants, trace minerals, or metals in the ground water table (Priyadarshi, 2001).

Farming: In a study conducted by the Henry Ford Health System, farming as an occupation was significantly associated with PD (OR, 2.79; 95% CI:1.03-7.55). The association of farming with PD was maintained after adjustment for occupational herbicide exposure and was of borderline significance after adjustment for occupational insecticide exposure (Gorell, Johnson, Rybicki, Peterson, & Richardson, 1998). These results suggest that the increased risk of developing PD through occupational exposure to farming is distinct and cannot be accounted for by pesticide, insecticide, and herbicide exposure,.

Race and Ethnicity: Studies that examine the relationship between race and PD have had contradictory findings. A study conducted in a culturally diverse community in New York City using a disease registry, age-adjusted PD prevalence rates were lower for blacks than for whites and Hispanics. Incidence rates were highest among black men, but they were otherwise comparable across the sex and ethnic groups. By ethnic group, the cumulative incidence was

higher for blacks than for whites and Hispanics, but more deaths occurred among incident black cases (Mayeux et al., 1995).

Another study conducted by Kaiser Permanente in Northern California found the Hispanics were at higher risk for developing PD than other ethnicities. The age- and gender-adjusted rate per 100,000 was highest among Hispanics (16.6, 95% CI: 12.0-21.3), followed by non-Hispanic Whites (13.6, 95% CI: 11.5-15.7), Asians (11.3, 95% CI: 7.2-15.3), and Blacks (10.2, 95% CI: 6.4-14.0) (Van Den Eeden, 2003).

Yet another study that used the Pennsylvania state Medicaid claims dataset from 1999-2003 found the 4-year cumulative incidence of PD was 45 per 100,000; 54 per 100,000 among whites, 23 per 100,000 among African-Americans and 40 per 100,000 among Latinos ($p < 0.0001$), corresponding to a relative risk (RR) of PD of 0.43 for African-Americans ($p < 0.0001$) compared with whites.

SES: Canadian research that used population and census data to assign PD cases to urban and rural income quintiles found that after adjusting for age and sex, average prevalence and incidence estimates were significantly higher for the lowest income quintile than the highest quintile. The annual rate of increase in the PD prevalence was significantly different for the lowest urban and rural income quintiles (Lix et al., 2010).

Conversely, risk factors have been identified in PD that confer decreased risk.

<u>Environmental Factors of Inverse Risk</u>

Coffee, Smoking and NSAIDs have been found, over the course of many decades and in multiple studies, to reduce the risk of developing PD (Checkoway et al., 2002; Samii, Etminan,

Wiens, & Jafari, 2009). Coffee and tobacco have been shown to be protective in Drosophila models of PD (Trinh et al., 2010).

GxE, Coffee, and PD

One of the primary reasons that GxE in PD gained momentum was the evidence for interaction that was identified earlier this decade. The 2011 study, on the NeuroGenetics Research Consortium (NGRC) dataset (the same data that are now publicly available and was used for this research), performed a genome-wide association and interaction study (GWAIS), testing each SNP's main-effect plus its interaction with coffee, adjusting for sex, age, and two ancestry. In GWAIS, the most significant signal came from rs4998386 and the neighboring SNPs in *GRIN2A*. *GRIN2A* encodes an NMDA-glutamate-receptor subunit and regulates excitatory neurotransmission in the brain. Achieving $p_{2df} = 10^{-6}$, *GRIN2A* surpassed all known PD susceptibility genes in significance in the GWAIS. In stratified GWAS, the *GRIN2A* signal was present in heavy coffee-drinkers (OR = 0.43; $p = 6\times10^{-7}$) but not in light coffee-drinkers. Imputation revealed a block of SNPs that achieved $p_{2df}<5\times10^{-8}$ in GWAIS, and OR = 0.41, $p = 3\times10^{-8}$ in heavy coffee-drinkers (Hamza et al., 2011). The results of this study have been independently replicated in other datasets and provided additional evidence indicating PD protective effects of coffee drinking/caffeine intake as well as the interaction with glutamate receptor genotypes. (Yamada-Fowler, Fredrikson, & Söderkvist, 2014).

These types of studies are important because they have the potential to identify susceptibility genes missed in GWAS, and identify genes that represent the entry point of environmental factors in modifying disease biology.

Side A Introduction and Rationale

Smoking

Cigarettes are a tobacco product. Although a cigarette has many chemicals and additives in both the tobacco and the filter, one of the key ingredients in tobacco is the stimulant and psycho-active chemical, nicotine. Nicotine is an alkaloid and constitutes approximately 0.6 - 3.0 % of tobacco by weight. It is capable of readily crossing the blood-brain barrier when inhaled (Le Houezec, 2003)

Smoking has been consistently shown to have an inverse association with risk of developing PD. Ever having smoked cigarettes was associated with a reduced risk of PD (odds ratio (OR) = 0.5, 95% CI: 0.4 -0.8). A stronger relation was found among current smokers (OR = 0.3, 95% CI: 0.1-0.7) than among ex-smokers (OR = 0.6, 95% CI: 0.4-0.9), and there was an inverse trend with pack-years smoked ($p < 0.001$) (Checkoway et al., 2002). In a meta-analysis that looked at 44 case-control and 4 cohort studies, the risk for PD in smokers compared with never smokers was 0.59 (95% CI: 0.54–0.63) for ever smokers, 0.80 (95% CI: 0.69–0.93) for past smokers, and 0.39 (95% CI: 0.32–0.47) for current smokers. The relative risk per 10 additional pack-years was 0.84 (95% CI: 0.81–0.88) in case–control studies and 0.78 (95% CI: 0.73–0.84) in cohort studies (Hernán, Takkouche, Caamaño-Isorna, & Gestal-Otero, 2002). Additional meta-analyses with 54 published epidemiological studies on smoking and PD risk found similar results. The risk estimates for current smokers, former smokers, and ever (current and former) smokers was 0.31 (95% CI: 0.25-0.38), 0.72 (95% CI: 0.63-0.83) and 0.55 (95% CI: 0.51-0.59), respectively (Kiyohara & Kusuhara, 2011b).

Evidence suggests that it is the nicotine, the active ingredient in cigarettes that confers the inverse risk/neuroprotective effect and protects against neuronal insults in PD (Quik, 2004;

Quik et al., 2006; Quik, Perez, & Bordia, 2012) Consumption of plants from the Solanaceae family (the family of plants to which nicotine belongs) have also shown to have the same inverse effect on risk, albeit to a smaller extent (Nielsen, Franklin, Longstreth, Swanson, & Checkoway, 2013). Nicotine is now being considered for development of therapeutic agents for the management of PD (Quik et al., 2012).

Exposure (Smoking) and Mechanisms of Action

The environmental exposure that is the subject of this research is smoking, specifically nicotine. Nicotine as a compound has been well-characterized for its effect on the human body. One of its main mechanisms of action is its binding to nicotinic receptors in the brain, but nicotine can also alter DNA methylation. DNA methylation maybe a mechanistic link between nicotine exposure and the development of PD, when nicotine is a known environmental risk factor.

<u>Environmental Exposure (Smoking) and Methylation</u>

Cigarette smoke is considered one of the most powerful environmental modifiers of DNA methylation (Breitling, Yang, Korn, Burwinkel, & Brenner, 2011). Smoking is known to be significantly associated with epigenome-wide global hypomethylation, as well as gene-specific hyper-methylation in the promoter region (Shigaki et al., 2012; Zhang et al., 2011). Cigarette smoking is thought to impact methylation in a number of ways.

1. Cigarette smoke may modulate methylation through DNA damage and subsequent recruitment of DNA methyltransferase enzymes (DNMT's) (Huang et al., 2013; Mortusewicz, Schermelleh, Walter, Cardoso, & Leonhardt, 2005; Smith & Hansch, 2000; Suter et al., 2011).

2. Cigarette smoke may also modulate DNA methylation through nicotine effects on gene expression (E. W. Lee & D'Alonzo, 1993). Nicotine binds to and activates the nicotinic acetylcholine receptors (present abundantly in the central and peripheral nervous systems) and increases intracellular calcium, which leads to downstream activation of cAMP response element-binding protein, a key transcription factor for many genes (Shen & Yakel, 2009). Possibly acting through this pathway, nicotine has been shown to downregulate DNMT1 mRNA and protein expression in mouse brain neurons (Satta et al., 2008).

3. Cigarette smoke may alter DNA methylation indirectly through the modulation of expression and activity of DNA-binding factors. It has been demonstrated that cigarette-smoke condensate increases Sp1 expression and binding to DNA in lung epithelial cells (Di, Zhao, & Harper, 2012; Mercer, Wallace, Brinckerhoff, & D'Armiento, 2009). Sp1 is a common transcription factor that binds to GC-rich motifs in gene promoters and plays a key role in early development; as such, it may prevent de novo methylation of CpGs within these motifs during early embryogenesis (Han, Lin, & Hsieh, 2001; Kadonaga, Carner, Masiarz, & Tjian, 1987).

4. Cigarette smoke may alter DNA methylation via hypoxia. Cigarette smoke contains carbon monoxide that binds to hemoglobin (competitively with oxygen) and thus decreases tissue oxygenation (Olson, 1984). Hypoxia, in turn, leads to the HIF-1α-dependent upregulation of methionine adenosyltransferase 2A, which is an enzyme that synthesizes S-adenosylmethionine, a major biological methyl donor critical for DNA methylation processes (Q. Liu et al., 2011).

Why are Causal Mechanisms Important in the GxE between Smoking and Genes in PD?

1. Smoking (and nicotine) is a high risk exposure for several diseases, including cancers, heart disease, and diabetes (Eyre, Kahn, & Robertson, 2004). However, its inverse risk in PD is well-established. This makes it challenging to advocate for nicotine-based neuroprotective therapies in PD, and simultaneously develop a public health message around nicotine/smoking and PD.

2. Nicotine has the potential to provide therapeutic targets in PD. It is critical to understand the modification (s) caused by nicotine that is causal in conferring decreased PD risk. These genes/pathways may be targeted by other therapeutics and will provide utility but without the deleterious effect of nicotine and advance this research towards translation and public health/clinical utility.

Rationale for GxE Analysis

PD is a multifactorial disease. Smoking is a well-established inverse-risk factor for PD, whose mechanism of neuroprotection has not been well-characterized. Researching GxE in PD (smoking-gene interactions) may lead to a better understanding of disease pathogenesis in PD. Identifying genes that are known to be differentially methylated by smoking provides a starting point to focus on a targeted set of SNPs and genes identified *a priori*. These SNPs (rSNPs) are hypothesized to interact with the environmental exposure in a manner that alters pre-existing methylation patterns. This provides a way to assess known risk factors by assigning the environmental exposure to altered biological function. This type of research has the potential to identify new susceptibility genes. Thus genes involved in PD, identified from an environmental perspective represent genes that are modifiable by an environmental exposure,

thus providing strong leverage points for research utility. Further, this more targeted approach of selecting genes also results in a reduction of statistical comparisons as compared to GWAS approaches.

Materials and Methods

Hypothesis

In identifying genes that are either differentially methylated by smoking or involved in metabolism of nicotine, we hypothesized that regulatory SNP's underlie a proportion of the regulatory differences in gene expression caused by methylation between cases and controls in PD, and an association analysis of SNPs in these genes could capture genetic differences between cases and controls in interaction with smoking.

Materials

Data Source

This research utilizes pre-existing and publicly available data, collected by the NeuroGenetics Research Consortium (NGRC). NGRC is a multi-center study of genetic and environmental risk factors in PD. The Consortium is currently comprised of eight movement disorder clinics in the four states (Oregon, Washington, Georgia, and New York). The genetic component of the study has been funded by the NIH since 1998. In 2004, the NGRC was initiated into the Michael J Fox Foundation as a Global Genetic Consortium funded by the Foundation. Created in 2004, the study's epidemiology arm introduced the environmental exposure questionnaire to the study (dbGaP, 2012). The NGRC study has a large number of samples in the dataset and has been the source of data for several publications thus far. The publicly available data, including genomic and environmental data were downloaded from dbGaP (dbGaP Study Accession: phs000196.v2.p1) (Mailman et al., 2007).

The NGRC Recruitment Process

All subjects were recruited at one of the eight NGRC-affiliated movement disorder clinics from among the four states. The study collected tissue (blood) samples from cases (n=2013) and controls (n=1995). Approximately 85 % of the cases and 85 % of the controls who were invited to volunteer in the study consented to be a part of the process (McCulloch et al., 2008). There was no preference in enrollment based on age of onset or family history of disease.

The subjects included in the GWAS were all white and self-identified as white-Americans or white- Europeans. This was done to minimize confounding caused by population substructure. In addition, subjects missing data on any of three covariates (site of recruitment, age (at blood draw) or sex), patients with onset before age 21, or patients with age of blood draw before age 20 were excluded from recruitment into the study. Both cases and controls consented to participate and donated blood for DNA extraction.

To qualify as a control in the study, a participant had to be genetically unrelated to the cases and to each other. Controls were spouses and community volunteers genetically unrelated to the patients. The controls did not fill out a standardized family history questionnaire, and only those recruited after 2004 filled out the Environmental Exposure Questionnaire (EEQ). The controls also had to be free of neurodegenerative disease by self-report or exam. Excluded diseases included Alzheimer's disease, bipolar disorder, multiple sclerosis, amyotrophic lateral sclerosis, ataxia, dystonia, Parkinson's disease, autism, dementia, epilepsy, stroke and schizophrenia.

To qualify as a case in the study, a participant had to have a confirmed diagnosed of PD by a neurologist, as diagnosed using the U.K Brain Bank Diagnostic Criteria(dbGaP, 2012). Cases were also required to be genetically unrelated to the controls and to each other. All cases filled out a standardized family history questionnaire, and those recruited after 2004 needed to also fill out the Environmental Exposure Questionnaire (EEQ).

The Diagnostic Process

Lewy bodies in the brain are the pathological hallmark of PD. However, there are no tests currently available to confirm the presence of the biomarker (Lewy body) in the Substantia nigra, short of an autopsy. This makes diagnosis of PD rather challenging. The U.K Parkinson's Disease Society Brain Bank Clinical Diagnostic Criteria is the gold standard for diagnosis of PD in the absence of any biomarker-based tools (Gibb,W.L, 1988). The criteria follow a three-step process.

Diagnostic Criteria

The first step evaluates whether the symptoms are Parkinsonian. The presence of bradykinesia (slowness of voluntary movement) along with at least one of three of the following symptoms: muscular rigidity, resting tremor and postural instability (not caused by loss of cerebellar, vestibular, visual dysfunction), are collectively used to make that diagnosis. The first step establishes whether the patient has symptoms of the Parkinsonian umbrella of diseases.

The second step is exclusionary, and excludes those diagnoses that have a non-PD etiology. This includes assessing the symptoms for other causative sources such as exposure to the neurotoxin precursor MPTP (1-methyl-4-phenyl-1,2,3,6,tetrahydropyridine), history of

stroke, head injury or encephalitis, dementia, cerebral tumors, Babinski's sign (signaling the presence of disease or lesions in the spinal cord and brain)and a negative response to levodopa treatment (if malabsorption has been excluded). The NGRC dataset modified this criteria so as not to exclude subjects with a positive family history of PD. The second step serves to eliminate all those patients whose symptoms have other etiologic sources.

The third step is the presence of supportive criteria for PD, and requires the manifestation of at least three of the following: unilateral onset, rest tremor, progressive disorder, persistent asymmetry affecting the side of onset most, responsive (70-100%) to levodopa, severe levodopa-induced chorea, levodopa response for 5 years or more, and a clinical course of these symptoms charted for over 10 years.

The stringency of the diagnostic criteria ensures that a diagnosis of PD is validated through multiple steps. Furthermore, the mean difference in the dataset between age at diagnosis and age at blood draw (8.36 yrs) added another layer to the accuracy of the diagnosis. In the unlikely event of a misdiagnosis, given the difference in mean years between the two variables, the misdiagnosis likely surfaced well before any participation in the study. Thus, all subjects classified as cases were most likely true cases. In a study that evaluated the accuracy of these diagnostic criteria in identifying idiopathic PD (IPD), the positive predictive value of the clinical diagnosis of IPD was extremely high, at 98.6% (72 out of 73 cases were correctly diagnosed) (Hughes, Daniel, Ben-Shlomo, & Lees, 2002).

Environmental Exposure

The Environmental Exposure Questionnaire (EEQ) was a standardized self-administered questionnaire. Environmental exposure information was first collected in 2004, with the advent

of the epidemiologic arm of the study. Collection of genetic samples, however, began earlier than the collection of exposure information. As a result, exposure assessment data is limited to those subjects enrolled after 2004, especially in Oregon (K. M. Powers et al., 2008). Overall, there is exposure information for about 80% of the cases and controls in the dataset.

For this analysis, we used the questionnaire information gathered on smoking. The questionnaire included three questions: smoking ever/never status (yes/no 100 cigarettes), smoking load (pack years) and smoking status (never/former/current). The smoking ever/never variable had the most complete data, and consequently was the variable chosen for this analysis.

Genotyping and Molecular Analysis

These data were genotyped at the Johns Hopkins University Center for Inherited Disease Research (CIDR) in Baltimore for the NGRC research consortium. Unamplified DNA was collected from whole blood at concentrations ≥ 50 ng/µl via standard methods of peripheral blood extraction. Genotyping was done using Illumina Human Onmi1 v1-0B SNP chip. The chip was customized for optimal tag SNP content and genome coverage. The chip provided genotyped information on 1,051,295 SNPs evenly spaced on the genome for a genome-wide scan of observed genotypes (McCulloch et al., 2008).

Methods

Framework for Selection of Genes and Variants for Testing

We selected genes (and variants) using a two-step process:

Materials and Methods

Step 1: Identifying genes differentially methylated by smoking

Step 2: Prioritize and assign tiers to selected genes for the statistical analysis.

Step 1

An exhaustive literature review revealed 36 genes found to be differentially methylated by smoking. In addition to these genes, three xenobiotic genes that are the primary metabolizers of nicotine (in the liver) - *CYP2A6, FMO3*, and *UGT1A4*, were added in, for a total of 39 genes. These genes were hypothesized *a priori* to be genes that influence PD risk because of their ability to be epigenetically modified by the environmental exposure (smoking) or because of their role in nicotine metabolism. SNPs from the following genes were chosen for the analysis.

Table 1: Genes known to be differentially methylated by smoking

Gene	Study Design	Study	Publication
DNMT1, DNMT3A DNMT3B, DNMT3L	Review of multiple studies	Reviewing multiple studies	Cigarette smoking and DNA methylation (K. W. K. Lee & Pausova, 2013)
CHRNA5	Review of multiple studies	Nicotinic receptor alpha-5 gene in multiple studies	Genetic Epidemiology of Smoking Behavior and Nicotine Dependence (Korhonen & Kaprio, 2011)
AHRR , GPR15	Case-control	Mononuclear cells from 111 African-American women	Smoking Associated DNA Methylation Changes in Peripheral Blood Mononuclear Cells from African American Women and Weighted Protein-Protein Interaction Networks (Dogan, 2014)
FLT1	Case-control	Epigenetic data from 22 participants	Coordinated DNA methylation and gene expression changes in smoker alveolar macrophages: specific effects on VEGF receptor 1 expression (Philibert et al., 2012)
HIVEP3, GNG12, GFI1 ALPPL2, CACNA1D AHRR, TIAM2, MYO1G CNTNAP2, ZC3H3 LRP5, PCDH9, RARA	Case-Control	Genome-wide differentially methylated CpGs of current compared to never smokers.	Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation (Zeilinger et al., 2013)

LINGO-3, F2RL3			
GNG12, GFI1,ALPPL2, PDZD2, VARS, MYO1G, C14orf43, F2RL3, TMEM51 AHHR	Cohort	SABRE Cohort	Differences in smoking associated DNA methylation patterns in South Asians and Europeans (Elliott et al., 2014)
LGALS4	Matched pair tumor and normal tissue from cases	Lung cancer	Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression (Selamat et al., 2012)
GFI1, CYP1A1, AHRR CYP1B1, C14orf43, NOTCH1, LRP5, CYTL1, F2RL3, PRSS23 ALPI, ALLPL2	Case-control	Expressed and hypomethylated in current smokers - EUROBATS Muther project	Genome-wide Association Scans Identify Differentially Methylated and Expressed Regions Related to Smoking in Adipose Tissue (Pei-Chien Tsai, 2014)
CHRNB4, TERT, CHRNA3	Matched pair from cases normal & tumor tissue	Lung cancer	Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of *CHRNB4* (Scherf et al., 2013)

Step 2

Rank the SNPs from step 1 (39 genes), and prioritize SNPs by selecting candidate genes based on function and prior evidence in PD. Genes that were known to be methylated in PD, and in gene families with strong biological plausibility were selected and prioritized as Tier 1 (3 genes). All other genes (36 genes) were designated as Tier 2 (See Table 2).

Tier 1: SNPs in *DNMT1, CNTNAP2,* and *LINGO3*

Tier 2: All SNPs in the remaining 36 genes

Table 2: Genes and SNPs by Tier

Genes By Tier		
Tier	N(Genes)	N(SNPs)
Tier 1	3	747
Tier 2	36	1534
Tiers combined	39	2281

The evidence for prioritizing Tier 1 genes

DNMT1 codes for the enzyme DNA methyl transferase1 which catalyzes the transfer of a methyl group to cytosine to form the methylated 5-methyl cytosine. *DNMT1*, the maintenance DNA methylation enzyme, is abundantly expressed in the adult brain and is mainly located in the nuclear compartment, where it has access to chromatin. Investigating the mechanisms underlying altered DNA methylation in PD and dementia with Lewy bodies (DLB) shows evidence of reduction of nuclear *DNMT1* levels in human postmortem brain samples from PD and DLB patients, as well as in the brains of α-synuclein transgenic mice models. Furthermore, sequestration of *DNMT1* in the cytoplasm results in global DNA hypomethylation in human and mouse brains, involving CpG islands upstream of *SNCA*, *SEPW1*, and *PRKAR2A* genes. The association of *DNMT1* and α-synuclein might mediate aberrant subcellular localization of *DNMT1*. These results could indicate a novel mechanism for epigenetic dysregulation in Lewy body diseases, which might underlie the decrease in DNA methylation reported for PD and DLB (Desplats et al., 2011).

CNTNAP2 codes for the enzyme Contactin-associated protein-like 2, a neuronal transmembrane protein member of the neurexin superfamily involved in neural-glia interactions and clustering of potassium channels in myelinated axons. *CNTNAP2* was found to

be one of 10 differentially methylated autosomal genes that co-varied in brain and blood DNA from PD patients. These data suggest that the detection of differential methylation events pertinent to PD pathology is feasible from blood samples, and reinforces the hypothesis that methylation of select genes may be a common risk factor for PD and age-associated brain changes (Masliah, Dumaop, Galasko, & Desplats, 2013).

LINGO3 codes for the Leucine-rich repeat -and Ig domain containing 3, which is a nogo protein that blocks axon regeneration following injury. Nogo proteins are one of the most potent neurite growth inhibitors in the CNS, functioning as negative regulators during development and serving as stabilizers of neuronal wiring in the adult brain (Marklund et al., 2007). Another member of the same gene family, *LINGO1,* has been identified as a novel target in PD. *LINGO1* expression is elevated in the substantia nigra of PD patients compared with age-matched controls and in animal models of PD after neurotoxic lesions. *LINGO1* expression is present in midbrain dopaminergic (DA) neurons in the human and rodent brain (Inoue et al., 2007).

Table 3: Location of gene and number of SNPs per gene

Gene	Chromosome	Build 36 Gene Location *	SNPs	Tier
TMEM51	1p36.21 (Plus)	chr1:15,342,816-15,429,459	45	Tier 2
FMO3	1q24.3 (Plus)	chr1:171,050,018-171,096,959	5	Tier 2
HIVEP3	1p34 (Minus)	chr1:41,738,271-42,167,083	140	Tier 2
GNG12	1p31.3 (Minus)	chr1:67,929,737-68,081,730	54	Tier 2
GFI1	1p22 (Minus)	chr1:92,702,906-92,734,216	12	Tier 2
ALPPL2	2q37 (Plus)	chr2:232,969,796-232,993,669	5	Tier 2
ALPI	2q37.1 (Plus)	chr2:233,019,077-233,042,986	8	Tier 2
UGT1A4	2q37 (Plus)	chr2:234,282,177-234,366,690	91	Tier 2
DNMT3A	2p23.3 (Minus)	chr2:25,299,349-25,428,278	31	Tier 2
CYP1B1	2p22.2 (Minus)	chr2:38,138,250-38,166,827	16	Tier 2
CACNA1D	3p14.3 (Plus)	chr3:53,494,071-53,831,532	123	Tier 2

GPR15	3q11.2-q13.1 (Plus)	chr3:99,723,568-99,744,650	10	Tier 2
CYTL1	4p16-p15 (Minus)	chr4:5,057,215-5,082,098	10	Tier 2
TERT	5p15.33 (Minus)	chr5:1,296,287-1,358,162	16	Tier 2
PDZD2	5p13.3 (Plus)	chr5:31,824,753-32,156,795	166	Tier 2
AHRR	5p15.3 (Plus)	chr5:347,292-501,405	46	Tier 2
TIAM2	6q25.2 (Plus)	chr6:155,443,115-155,630,549	94	Tier 2
VARS	6p21.3 (Minus)	chr6:31,843,276-31,881,691	18	Tier 2
CNTNAP2	7q35 (Plus)	chr7:145,434,386-147,759,019	721	Tier 1
MYO1G	7p13-p11.2 (Minus)	chr7:44,958,786-44,995,193	9	Tier 2
ZC3H3	8q24.3 (Minus)	chr8:144,580,968-144,704,763	42	Tier 2
NOTCH1	9q34.3 (Minus)	chr9:138,498,717-138,570,059	27	Tier 2
LRP5	11q13.4 (Plus)	chr11:67,826,684-67,983,319	45	Tier 2
PRSS23	11q14.1 (Plus)	chr11:86,179,139-86,209,921	11	Tier 2
FLT1	13q12 (Minus)	chr13:27,764,389-27,977,265	66	Tier 2
PCDH9	13q21.32 (Minus)	chr13:65,764,967-66,712,464	240	Tier 2
C14orf43	14q12 (Plus)	chr14:73,241,578-73,333,649	30	Tier 2
CYP1A1	15q24.1 (Minus)	chr15:72,788,936-72,814,930	9	Tier 2
CHRNA5	15q24 (Plus)	chr15:76,634,961-76,683,515	92**	Tier 2
CHRNA3	15q24 (Minus)	chr15:76,664,706-76,710,377	"**	Tier 2
CHRNB4	15q24 (Minus)	chr15:76,693,691-76,730,642	"**	Tier 2
RARA	17q21 (Plus)	chr17:35,708,972-35,777,420	13	Tier 2
DNMT1	19p13.2 (Minus)	chr19:10,095,022-10,176,811	12	Tier 1
F2RL3	19p12 (Plus)	chr19:16,850,826-16,873,830	11	Tier 2
LINGO3	19p13.3 (Minus)	chr19:2,230,774-2,269,156	14	Tier 1
LGALS4	19q13.2 (Minus)	chr19:43,974,155-44,005,422	13	Tier 2
CYP2A6	19q13.2 (Minus)	chr19:46,031,283-46,058,192	3	Tier 2
DNMT3B	20q11.21 (Plus)	chr20:30,803,852-30,870,823	21	Tier 2
DNMT3L	21q22.3 (Minus)	chr21:44,483,297-44,513,881	12	Tier 2
		TOTAL SNPs	2281	

* Includes SNPs on loci 10kb upstream and 10kb downstream from gene location
** the 3 genes *CHRNA3, CHRNA5,* and *CHRNB4* overlap and the entire locus contributed 92 SNPs to the analysis

Selection of SNPs for the Analysis

For the purpose of this analysis, a subset of the SNPs genotyped as part of the NGRC study using the Illumina Human Onmi1 v1-0B SNP chip were selected. These SNPs were localized to the subset of genes found to be differentially methylated by smoking as described above. The gene locus for each gene included regions 10kb upstream and downstream of each

gene, to capture any potential SNPs on 5' and 3' regulatory elements. A total of 2281 SNPs were chosen for the analysis representing the 39 genes found to be differentially methylated by smoking. The SNPs were genotyped using build NCBI36/hg18. Because maximizing power and reducing the number of multiple comparisons were priorities, and because the imputed SNPs collectively represent the same set of signals as the tag SNPs genotyped, imputed data were not used to expand the number of SNPs for this analysis. The genotyped SNPs were all common SNPs with minor allele frequencies (MAF) above 1%.

In all there were 2281 SNPs across 17 chromosomes in the 39 genes selected *a priori* (see Table 3). The 2281 SNPs were assigned to 2 groups to prioritize genes for this analysis. Tier 1 comprised 3 genes (*CNTNAP2, DNMT1,* and *LINGO3)* and 747 SNPs. The other thirty six genes containing 1534 SNPs comprised Tier 2. The number of tag SNPs per gene varied from 3 SNPs (*CYP2A6*) to 721 SNPs (*CNTNAP2*), depending on the size of the gene, and selection of SNPs chosen as tag SNPs for the SNP chip.

The optimal set of SNPs to analyze based on my hypothesis of rSNPs would have been SNPs chosen on genes differentially methylated by smoking and then prioritized by SNPs in CpG islands within genes (SNPs most likely to exhibit regulatory function based on their location). However, since we were working with pre-existing data, collected on an earlier SNP chip that was not densely genotyped, all available SNPs from the list of genes chosen were selected for the analysis.

Materials and Methods

Quality Control

Genotyped SNPs were processed for quality control checks on SNPs for missingness, duplicates, minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) and imputation quality.

Statistical Analyses

STATA version 12 was the software used for most of the data analysis (StataCorp., 2011). The two datasets: the genotyped reference panel information (with SNPs from the 39 genes) and covariate information were combined. Since the genotyped data was initially obtained on NCBI Build 36.3, all chromosomal locations map to that build. However, in order to standardize the genomic location of the SNPs, all SNPs that were significant were converted to Build 37.3 and all genes and locations mapped to the SNPs use the newer build using UCSC's Genome Browser (Kent et al., 2002).

Logistic regression was used in this analysis. It was well-suited to evaluate potential GxE interactions and risk of PD and it also allowed for adjustment of potentially confounding factors such as age at blood draw, and sex. Specifically, the model included age at blood draw, sex, an additive SNP effect, the binary environmental factor smoking (ever/never), and an interaction term (environmental factor*SNP). The adjustments allow for minimizing the underlying variability in the data as well as adjust imbalanced baseline variables known to be related to the outcome.

SimpleM (R package) was used to generate Meff values for multiple comparisons generated by the association analysis. The Bonferroni method of adjusting for multiple comparisons is easy to compute, but is well known to be conservative in the presence of LD.

Meff is efficient and accurate and an appropriate choice for multiple testing adjustment when there is high intermarker LD in the SNP data set (Gao, Starmer, & Martin, 2008; R Foundation for Statistical Computing, 2008). After considering LD between SNPs as calculated by SimpleM, there were 398 effectively independent tests for Tier 1, and 881 effectively independent tests for Tier 2. A correction was then applied using the Meff value of number of effective independent tests and significance threshold values for each Tier were calculated. They were for Tiers 1 and 2 ($p=1.26 \times 10^{-4}$ and $p=5.67 \times 10^{-5}$), respectively (see Table 4).

Table 4: Significance threshold

Tier	# SNPs per Tier	Meff value	α/n	Significance threshold
1	747	398	.05/398	1.26×10^{-4}
2	1534	881	.05/881	5.67×10^{-5}

The regression models were run for Tiers 1 and 2 separately and interaction p-values were generated. The data were modeled to evaluate evidence for interactions between smoking ever/never and the SNPs chosen for the analysis. Interactions models were run for pairwise analysis GxE (SNP*smoking) and the results were evaluated with respect to the risk they conferred on PD susceptibility. Interaction odds ratios and p-values were computed to assess the strength of the association.

Results

The descriptive demographics (Table 5) reveal that the mean age at blood draw was greater in controls than in cases. This was intentionally done during data collection, in order to eliminate the possibility that some of the controls could be pre-symptomatic cases if their age at blood draw was less than that of the cases. The proportion of subjects in the ever/never category for smoking was similar between cases and controls.

Table 5: Descriptive demographics of the NeuroGenetics Research Consortium (NGRC) dataset

Descriptive Demographics				
Covariates	NTotal (Ncases, Ncontrols)	Total (N=3986)	Cases (N=2000)	Controls (N=1986)
Age onset (mean, sd)	1999 (1999, 0)	58.39, 11.93	58.39, 11.93	n/a
Age onset (p25, median, p75)	1999 (1999,0)	50, 58.39, 67	50, 58.39, 67	n/a
Age diagnosis (mean, sd)	1950 (1950, 0)	60.42, 11.40	60.42, 11.40	n/a
Age diagnosis (p25, median, p75)	1950 (1950, 0)	53, 60.42, 69	53, 60.42, 69	n/a
Age blood draw (mean, sd)	3986 (2000, 1986)	68.78, 12.59	67.26 , 10.67	70.32 , 14.09
Age blood draw (p25, median, p75)	3986 (2000, 1986)	60, 69, 78	60, 68.5, 75	60, 71, 83
Age at questionnaire (mean, sd)	3120 (1611, 1509)	70.57, 12.16	68.53, 10.71	72.75, 13.2
Age at questionnaire (p25, median, p75)	3120(1611, 1509)	62.3, 70.57, 80.1	61.5, 68.58, 76.3	63.5, 72.25, 85.2
Sex (N male % male)	2115 (1346, 769)	53.06%	67.30%	38.72%
Sex (N female, % female)	1871 (654, 1217)	46.94%	32.70%	61.28%
Smoking (N ever_never, % ever)	3104 (1598, 1506)	46.00%	45.81%	46.22%
Smoking (N ever_never, % never)	3104 (1598, 1506)	54.00%	54.19%	53.78%

Results

All 2281 SNPs chosen for the analysis cleared the quality control process, and none were dropped. Logistic regression was conducted on the genes from both tiers and the results from the logistic regression models from both tiers showed no significant evidence for interaction for the environmental factor (smoking ever/never) with PD in any of the regions evaluated, after adjusting for multiple comparisons. Shown in Tables 6 & 7 are the results for each of the Tiers presented in each analysis (using smoking ever/never variable). The threshold for statistical significance after adjusting for multiple comparisons calculated using SimpleM's Meff values were Tier1 p-value=1.25 x 10-4 (.05/398, the equivalent of α=0.05 after multiple comparisons) and Tier 2 p-value= 5.67 x 10-5 (.05/881, the equivalent of α=0.05 after multiple comparisons). The p-values and odds ratios in these results are interaction p-values and interaction odds ratios.

Table 6: Tier1 smoking analysis (adjusted for age at blood draw, sex, smoking (ever/never 100 cigarettes))

Tier 1 Interaction									
SNP	N	OR	95% CI (lower)	95% CI (upper)	Z	p-val	Gene	Position (Build 37)	Location
rs2101777 (A/G)	3099	1.89	1.14	3.13	2.46	0.013873	*CNTNAP2*	7:146699636	Intron
rs10263021 (C/T)	3104	1.33	1.06	1.68	2.42	0.015343	*CNTNAP2*	7:148228400	Intron
rs12216694 (A/G)	3104	2.33	1.12	4.83	2.27	0.023215	*CNTNAP2*	7:146407539	Intron
rs10248899 (A/C)	3100	1.70	1.07	2.70	2.23	0.025842	*CNTNAP2*	7:146341974	Intron
rs17507650 (C/T)	3104	0.65	0.44	0.96	-2.16	0.031081	*CNTNAP2*	7:147468405	Intron
rs1015932 (A/G)	3102	0.78	0.62	0.98	-2.13	0.033014	*CNTNAP2*	7:147494245	Intron
rs6964680 (G/T)	3104	1.46	1.03	2.08	2.12	0.03392	*CNTNAP2*	7:148279947	Intron
rs7810054 (A/G)	3102	2.39	1.06	5.35	2.11	0.034885	*CNTNAP2*	7:146294039	Intron
rs6953901 (A/G)	3083	1.97	1.04	3.75	2.07	0.038848	*CNTNAP2*	7:148308151	Intron
rs971818 (C/T)	3104	0.79	0.62	0.99	-2.05	0.039906	*CNTNAP2*	7:147506105	Intron

In the Tier 1 interaction analysis, of the three genes analyzed (*CNTNAP2, DNMT1,* and *LINGO3*), the variant that displayed the strongest evidence for interaction with smoking was a non-

coding polymorphism rs2101777 (A/G), an intronic SNP on the *CNTNAP2* gene(OR = 1.88, 95% CI: 1.13- 3.12, *p* = 0.013873). However, the p-value did not rise to the pre-determined level of significance (p=1.25 x 10-4, the equivalent of α=.05 after multiple comparisons). The SNPs with the most extreme p-values in Tier 1 had a p-value range of p=0.013873 to p=0.039906. Only *CNTNAP2* was represented among the extreme p-values, but *CNTNAP2* also contributed 721SNPs out of the 747 SNPs in Tier1.

Table 7: Tier 2 smoking analysis (adjusted for age at blood draw, sex, smoking (ever/never 100 cigarettes))

Tier 2 Interaction									
SNP	N	OR	95% CI (lower)	95% CI (upper)	Z	p-value	Gene	Position (Build 37)	Location
rs349430 (A/G)	3102	0.62	0.47	0.82	-3.43	0.000613	*HIVEP3*	1:41776488	Intron
rs11580042 (C/T)	3103	0.69	0.56	0.86	-3.41	0.000647	*FMO3*	1:172841666	Intergenic
rs2612026 (A/G)	3104	1.61	1.21	2.14	3.29	0.000997	*CACNA1D*	3:53742968	Intron
rs898419 (G/T)	3104	1.59	1.20	2.11	3.23	0.001251	*CACNA1D*	3:53749758	Intron
rs7373113 (A/C)	3103	1.59	1.20	2.13	3.18	0.001496	*CACNA1D*	3:53756027	Intron
rs1325432 (A/C)	3104	0.63	0.47	0.84	-3.13	0.001721	*GFI1*	1:92475293	3" UTR
rs1020819 (G/T)	3104	0.65	0.49	0.87	-2.94	0.003289	*CACNA1D*	3:53728686	Intron
rs1020820 (C/T)	3104	1.56	1.16	2.10	2.91	0.003648	*CACNA1D*	3:53728542	Intron
rs3774430 (A/G)	3103	2.58	1.35	4.93	2.87	0.004081	*CACNA1D*	3:53532423	Intron

The results of the Tier 2 interaction are summarized in Table 7. Of the 36 genes analyzed, the variant that displayed the strongest evidence for interaction with smoking was a non-coding polymorphism rs349430 (A/G), an intronic SNP of *HIVEP3* (OR = 0.62, 95% CI: 0.47-0.81, *p* = 0.000613). Although the p-values were, on average, much smaller in Tier 2, the p-values in Tier 2 still did not rise to the pre-determined level of significance (p=5.67 x 10-5, the equivalent of α=0.05 after multiple comparisons). The SNPs with the most extreme p-values in

Tier 2 had a p-value range of p=0.000613 to p=0.004081. Of the 36 genes analyzed in Tier 2, five genes were represented among the top 10 SNPs (*HIVEP3, GFI1, NOTCH1, CACNA1D, FMO3)*.

Discussion

To summarize, this research sought to identify PD susceptibility genes that confer altered risk through interactions with smoking, using SNP data on thirty nine genes in a tiered two-stage design. We selected 39 genes, using a novel gene selection method: genes found to be differentially methylated by smoking (36 genes) and xenobiotic genes encoding enzymes that metabolize of nicotine (3 genes), and hypothesized *a priori* to be genes that influence PD risk on account of their ability to be epigenetically modified by the environmental exposure (smoking). In this set of 39 genes, 2281 SNPs were analyzed for GxE with smoking. None of the p-values were significant after accounting for multiple testing. The most extreme p-value in the first tier (Tier 1) of the analysis (p = 0.013873 on *CNTNAP2*) did not reach the pre-determined level of significance (p=1.25 x 10-4). The most extreme p-value in the second tier (Tier 2) of the analysis (p = 0.000613 on *HIVEP3*) did not reach the pre-determined level of significance (p=5.67 x 10-5). A point of consideration in interpreting these results observed is that these are not p-values for main effect but p-values for interaction effect. Thus, findings for this analysis represent that while interactions were detected, the interactions were not statistically significant, but are suggestive.

Missing Heritability of Complex Diseases and GxE

The missing heritability of complex diseases is well documented. Multiple avenues of research are being pursued to account for the missing heritability, however, most continue to focus on

finding the genetic variants, which also includes the genetic markers of regulation (Eichler et al., 2010; Manolio et al., 2009). In these geno-centric approaches, the population is assumed to share a common environment, or environment is adjusted for, or in some cases ignored. However, if environmental factors modify the risk of disease (through interactions with genes), it is possible that important genomic effects will be missed. GxE interactions are likely salient in both architecture of complex diseases and the rising prevalence of chronic diseases, and they may account for some of the missing heritability, but they will not appear in the heritability estimates. Genetic architecture may be thought of as the genetic basis of phenotypic traits. Genetic architecture required for biological function is complex, and involves not only genetic variation, but also the regulation and orchestration of multiple genes that combine and act in concert along various pathways to produce the phenotype. Understanding complex, multifactorial diseases require the modeling of this genetic architecture, which includes modeling the impact of environmental exposures in interaction with genes. Accounting for these processes may lead to new insights regarding complex diseases, including PD.

GxE Limitations

Overall, this analysis did not observe statistically significant evidence of interaction for any of the genes hypothesized to be in interaction (Tables 7 & 8) with PD risk. There are several possible explanations for the failure to find statistically significant evidence. These limitations apply to both the genotype-related data as well as the exposure-related data.

<u>Genotype-Related Limitations</u>

With most genetic association studies and GxE studies in particular, there are a number of well-documented limitations.

Statistical Power

Lack of statistical power is common to all genome-wide association studies, and is more pronounced in GxE research. While the novel method of gene selection reduced the number of multiple comparisons, and the two- stage design to prioritize the genes reduced the number of multiple comparisons and increased power, it may not have been enough to detect effect at a level of statistical significance.

Interaction p-values

Identifying interaction p-values at levels of statistical significance required for GWAS is a challenging hurdle to overcome. GxE models are multiplicative and for the interaction p-values to reach significance thresholds that are the same order of magnitude as models of main effect in GWAS requires the presence of very strong interaction effects. While such effects may well exist, the absence of such strong effects is neither evidence of lack of interaction between the environmental exposure examined and a SNP at a given locus, nor is it evidence of lack of biological plausibility. It is however indicative of model limitation. With the methods that we currently possess, we are unable to detect interactions in a manner that will appropriately weight the significance of the interaction to disease phenotype. Additionally, if GxE interactions are strongest for rare alleles, this will be very hard to detect using sample sizes and SNP data available in most case-control or cohort studies.

SNP Selection

The dataset used for this analysis (NGRC) contains genotyped information on 1,051,295 SNPs evenly spaced on the genome for a genome-wide scan of observed genotypes. To the extent that the genotyped tag SNPs are not in LD with the regulatory SNPs, there may be misclassification in these data because the tag SNP signals may not be capturing signals from the regulatory variants involved in gene regulation through smoking interactions in these genes. Imputing data from the tag-SNPs would have yielded much the same results because the imputed SNPs are those in LD with the tag SNPs, and likely tag the same signals. The results from this analysis may have produced stronger signals if we were able to map tag SNPs on to CpG islands and rank the SNPs for analysis based on their location on known regulatory regions.

Exposure Data-Related Limitations

Misclassification of environmental variables is also of concern in GxE studies and can also reduce power.

Accuracy of exposure measurement

Environmental exposure questionnaires record people's retrospective reports of exposure. Retrospective recall can be subject to recall bias, and in cases of neurodegenerative diseases, it could also include biases stemming from cognitive decline. Recall bias has been found to be of particular concern in case-control studies with retrospective aspects (such as exposure evaluation) and evaluating disease etiology.

Smoking variables in the data and methylation patterns

The environmental exposure questionnaire gathered information in the three questions related to smoking were: smoking ever/never (whether or not a subject has smoked at least a 100 cigarettes or more), smoking status (never/ former/current) and smoking load (a

continuous variable denoting pack years). Smoking ever/never had the least missing data and was chosen as the variable to analyze. However, when hypothesizing about DNA methylation and subsequent alteration of gene expression as an effect of smoking, research shows that depending on cessation time and pack-years, methylation levels in former smokers were found to be very similar to the patterns seen in never smokers (Zeilinger et al., 2013). Using smoking ever/never as a variable, neither captures current smoking patterns nor the smoking load. More detailed exposure data will help to understand methylation patterns altered by smoking and PD risk.

Temporality

If smoking is (through altered patterns of methylation) associated with PD etiology, it is important to map smoking start/stop chronology vis-a-vis onset of disease, and the environmental exposure data did not contain information on when the subject started smoking or stopped smoking, or if they were smoking before/during/after disease onset. To add to these challenges, among cases, blood drawn for genotyping were drawn decades after disease onset in some instances. Moreover, levels of nicotine and/or metabolites were not evaluated from blood samples when blood was drawn. The inability of these data to provide estimates of temporality between disease onset and exposure are an important limitation to the analysis.

These limitations speak to the need for the development and integration of new tools and methods that are designed specifically to maximize the possibility of detecting interactions from statistical (significance and effect sizes), biological (functional consequence), and environmental (accurate biomarkers of exposure) perspectives.

Additional Analyses

These were conducted to try and address the accuracy of environmental exposure and assign function to the SNPs from the results of the statistical analysis.

Sensitivity Analysis

To evaluate the potential impact of temporality, a sensitivity analysis was conducted on the data. The intent here was to try to accurately stratify the data and create a subset of cases among whom DNA methylation levels (in blood) were as close as possible to levels at disease onset, and therefore rSNP signals would be assumed to be closely allied to methylation patterns. A new variable was created: 'methylated', defined among cases as [age of blood draw - age of disease onset <=10 years]. This eliminated 637 cases (those among whom age at blood draw was more than 10 years after age at onset of PD). A total of 3349 subjects comprised the N for the sensitivity analysis, 1363 cases and 1986 controls. The results generated from this analysis were comparable to the previous analysis and did not reach significance (Tables in appendix). The most extreme p-value in the first tier (Tier 1) of the sensitivity analysis (p = 0.011377) was on the *CNTNAP2* gene. The most extreme p-value in the second tier (Tier 2) of the sensitivity analysis (p = 0.000442768) was on the *CACNA1D* gene.

Functional Analysis

In thinking through GxE and its connection with rSNPs, it is evident that statistical significance cannot be the only measure used to corroborate interactions. To find evidence for regulation, these data also need to be characterized functionally. While annotation and functional characterization of SNP signals now have extensive databases, epigenetics is still a

nascent field when compared to genomics. Therefore it made sense to choose the interaction association analysis as the primary measure of data investigation and then apply a bioinformatics filter to characterize function to the results of the main analysis (Tiers 1 and 2). This way, we maintained the rigor of the statistical analysis while using bioinformatics to elucidate regulatory signals in these SNPs. The top 100 SNPs from each Tier were evaluated for evidence of function using the Regulome DB database. The database assigns a score (from 1-6) based on various lines of evidence (eQTL, TF binding, TF motif, DNase footprint, DNase peak). Score values in the range from 1a-1f represent the most confidence that a variant lies in a functional location and likely results in a functional consequence (Boyle et al., 2012). From the top hundred SNPs in each Tier of the regression analysis, we annotated the variants that had scores of 1a-1f (and known to be functionally significant and associated with differences in gene expression).

Annotation Results

Table 8: Tier 1 functional analysis

Functional Analysis Tier 1				
Coordinate	SNP	Score	Interpreting Score	Gene
chr7: 147973956	rs17170936	1f	eQTL + TF binding / DNase peak	*CNTNAP2*

Table 9: Tier 2 functional analysis

Functional Analysis Tier 2				
Coordinate	SNP	Score	Interpreting Score	Gene
chr17:38460374	rs2120200	1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak	*RARA*
chr17: 38460860	rs36030243	1d	eQTL + TF binding + any motif + DNase peak	*RARA*

Tier 1 Gene CNTNAP2 and Annotated Results

One SNP on the Tier 1 gene *CNTNAP2* had functional scores suggesting strong evidence of functional consequence. *CNTNAP2* on chromosome 7 is the largest gene in this analysis, spans 2,304,633 bases including SNPs 10,000 bp upstream and downstream of the gene. It contributed 721 SNPs to the analysis. The gene codes for a Contactin-associated protein-like 2, a neuronal transmembrane protein member of the neurexin superfamily involved in neural-glia interactions and clustering of potassium channels in myelinated axons. *CNTNAP2* was found to be one of 10 differentially methylated autosomal genes that co-varied in brain and blood DNA from PD patients (Masliah et al., 2013).

Tier 2 Gene RARA and Annotated Results

The Tier 2 gene that emerged with significant evidence of function is the *RARA* gene on chromosome 17, spanning 48,448 bases including SNPs 10,000 base pairs upstream and downstream of the gene. The *RARA* gene codes for retinoic acid receptor alpha. The protein, RARα, regulates transcription in a ligand-dependent manner and has been implicated in regulation of development, differentiation, apoptosis, granulopoeisis, and transcription of clock genes. *RARA* is known to be differentially methylated by smoking (Zeilinger et al., 2013). Retinoic acid receptors are a part of the RA pathway. Retinoic acid (RA) signaling pathway. RA is involved in the induction of neural differentiation, motor axon outgrowth and neural patterning. Elevated RA signaling in the adult triggers axon outgrowth and, consequently, nerve regeneration. RA is also involved in the maintenance of the differentiated state of adult neurons, and disruption of RA signaling in the adult leads to the degeneration of motor neurons (motor neuron disease), the development of Alzheimer's disease and, possibly, the

development of Parkinson's disease (Maden, 2007).

LD in *RARA* SNPs from Functional Analysis and Statistical Analysis

Of the 13 SNPs in RARA, two SNPs showing evidence for functional consequences, (rs2120200 and rs36030243) are SNPs that alter gene expression and bind transcription factors. The Tier 2 sensitivity analysis also had two SNPs from the same gene among the top 10 SNPs (rs12103711 and rs2715553). Rs12103711, an intronic SNP of RARA (OR = 1.80, 95% CI: 1.23-2.65, p = 0.002802) and rs2715553, an cSNP of RARA (OR = 0.72, 95% CI: 0.57 - 0.91, p = 0.00563). We evaluated LD between the four SNPs in RARA to determine if these SNPs were all pinpointing a single signal with evidence for regulation. We found that the four SNPs are in strong LD with each other, although the rs2715553, a cSNP is physically some distance away (35 kb) from the other three SNPs. We found rs2120200 and rs36030243 are rSNPs on the RARA gene that interact with smoking and are in LD with a coding SNP on the same gene.

We performed the same analysis for the Tier 1 gene *CNTNAP2* and none of the top 10 SNPs from the regression models were in LD with each other.

Table 10: LD among the four SNPs in *RARA*

SNP	Proxy	Distance	r^2	D'	Chromosome	Coordinate_HG18
rs2120200	rs36030243	486	1.000	1.000	chr17	35714387
rs2120200	rs12103711	1265	0.818	1.000	chr17	35715166
rs2120200	rs2715553	35945	0.082	1.000	chr17	35749846
rs36030243	rs12103711	779	0.818	1.000	chr17	35715166
rs36030243	rs2715553	35459	0.082	1.000	chr17	35749846
rs12103711	rs2715553	34680	0.101	1.000	chr17	35749846

Interpreting the LD Values

The four SNPs (rs2120200, rs36030243, rs1203711, and cSNP rs2715553) are in strong LD with each other as evidenced by their D' values. This indicates that there is no evidence for recombination with that block, but the low r^2 values between cSNP rs2715553 with the other three SNPs on the gene is evidence that none of the other three SNP are proxies for the cSNP. The low r2 values may be attributable to the difference in allele frequencies between rs2715553 as compared to the other three SNPs. rs2715553 has a minor allele frequency (MAF) of 0.4050, while the other three SNPs have MAF's in the range between 0.2780 - 0.2947. Although the cSNP is almost 36 kb away from the other three intronic SNPs, this is within the range of 60kb LD blocks, on average, for Northern -European populations (Reich et al., 2001).

The scope of this analysis requires measuring non-random association of alleles in these four SNPs on *RARA*, and these data are evidence of strong disequilibrium. While the regulatory SNPs have r^2 values over 0.8 with each other, the low r^2 value with the cSNP does not negate the fact that they are all co-inherited and in the same LD block. To clarify why the D' value is adequate to prove LD in this instance, the hypothesis for rSNP hinges on SNPs in regulatory regions whose alleles impact epigenetic states of modification which in turn impacts gene expression. The two rSNPs of functional consequence that we identified using bioinformatics on the *RARA* gene in this analysis are in LD with the coding SNP identified using logistic regression on the same gene. This is consistent with the rSNP hypothesis that non-coding variants in regulatory regions are statistically significant (or suggestive) in an association analysis, are located on non-protein coding regions of genes, have functional consequences, and are

associated with disease phenotypes by altering gene expression. While both SNPS are categorized as being likely to affect binding and linked to expression of a gene target, rs2120200 is categorized an enhancer and rs36030243 is categorized as weak transcription. The two SNPs (rs2120200 and rs36030243) could be cis-acting regulatory variants on the *RARA* gene. Cis-acting variants are non-coding variants that regulate the expression of the gene, and are in the vicinity of the gene they regulate. Finally, both functional rSNPs have been shown to be involved in epigenetic modifications (rs2120200 as an 'enhancer', and rs36030243 as 'weak transcription') in the substantia nigra of the brain, the brain tissue that is the hallmark of neurodegeneration in PD (see Figures in Appendix).

RARA - A Susceptibility Gene

Based on the strong functional evidence and suggestive statistical evidence, *RARA* could be considered a susceptibility gene in PD. It is a gene that is modified by smoking, and SNPs in *RARA* could potentially alter PD risk in interactions with smoking. There is need for more research to confirm that *RARA* plays a role in PD etiology through interactions with smoking, and subsequently the need for more research to assess potential therapeutic value in PD, especially in light of the fact that nicotine use does not lend itself to a public health message. Retinoic acid and the three genes that code for the retinoic acid receptor (*RARA, RARB*, and *RARG*) could be considered additional targets for research interest in PD. To our knowledge, *RARA* has not been evaluated specifically as a risk susceptibility gene in PD, although it has been evaluated as a susceptibility gene in other neurodegenerative diseases (Maden, 2007).

Next Steps

Some potential next steps to move this research forward include evaluating allele-specific effects on the four SNPs (from the functional analysis and sensitivity analysis) on *RARA*, and functionally evaluating all SNPs on *RARA* (from the NGRC data). Other steps include identifying CpG SNPs on *RARA* and performing GxE using those genotypes and validating these results in other PD datasets. Finally, it may help gain a better understanding of the RA signaling pathway and its effect on PD, if we also evaluate smoking interactions in the other two genes that code for the retinoic acid receptor (*RARB* and *RARG*).

In Conclusion

This analysis set out to conduct an exploratory analysis: that selecting genes by following the effect of the environmental exposure is a valid method of identifying susceptibility genes in GxE. This is also a socially-just method of conducting genomics research. In the absence of genetic data on vulnerable groups, researching environmental risk factors of inverse risk (on a largely white urban population) is one way by which PD research can be equitable. This method required us to alter how we usually investigate genetics - by not starting out with genes impacting the disease phenotype, but focus instead on the environment. We followed the effects of the environmental exposure into the body in order to pinpoint genes and rSNPs at the intersection between the environmental exposure and disease pathways. A key concept is that these genes are selected based on their interaction with smoking and

contribute to regulatory effects. We have used a novel approach to identifying SNPs and genes that warrant further study in understanding the interaction between smoking and PD.

References

1. Baylin, S. B. (2005). DNA methylation and gene silencing in cancer. *Nature Clinical Practice Oncology*, *2*, S4–S11. http://doi.org/10.1038/ncponc0354

2. Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., … Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, *12*(1), R10. http://doi.org/10.1186/gb-2011-12-1-r10

3. Bjornsson, H. T. (2008). Intra-individual Change Over Time in DNA Methylation With Familial Clustering. *JAMA*, *299*(24), 2877. http://doi.org/10.1001/jama.299.24.2877

4. Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., … Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. http://doi.org/10.1101/gr.137323.112

5. Braveman, P. (2006). HEALTH DISPARITIES AND HEALTH EQUITY: Concepts and Measurement. *Annual Review of Public Health*, *27*(1), 167–194. http://doi.org/10.1146/annurev.publhealth.27.021405.102103

6. Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., & Brenner, H. (2011). Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. *The American Journal of Human Genetics*, *88*(4), 450–457. http://doi.org/10.1016/j.ajhg.2011.03.003

7. Bridget C. Booske, Jessica K. Athens, David A. Kindig, Hyojun Park, & Patrick L. Remington. (2010). *DIFFERENT PERSPECTIVES FOR ASSIGNING WEIGHTS TO DETERMINANTS OF HEALTH* (COUNTY HEALTH RANKINGS WORKING PAPER) (p. 22). Wisconsin: University of Wisconsin. Retrieved from https://uwphi.pophealth.wisc.edu/publications/other/different-perspectives-for-assigning-weights-to-determinants-of-health.pdf

8. Chang, V. W., & Lauderdale, D. S. (2009). Fundamental Cause Theory, Technological Innovation, and Health Disparities: The Case of Cholesterol in the Era of Statins. *Journal of Health and Social Behavior*, *50*(3), 245–260. http://doi.org/10.1177/002214650905000301

9. Checkoway, H., Powers, K., Smith-Weller, T., Franklin, G. M., Longstreth, W. T., & Swanson, P. D. (2002). Parkinson's Disease Risks Associated with Cigarette Smoking, Alcohol Consumption, and Caffeine Intake. *American Journal of Epidemiology*, *155*(8), 732–738. http://doi.org/10.1093/aje/155.8.732
10. Costa, J., Lunet, N., Santos, C., Santos, J., & Vaz-Carneiro, A. (2010). Caffeine Exposure and the Risk of Parkinson's Disease: A Systematic Review and Meta-Analysis of Observational Studiess. *Journal of Alzheimer's Disease*, (S1), 221–238. http://doi.org/10.3233/JAD-2010-091525

11. Coughlin, S. S. (1990). Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology*, *43*(1), 87–91. http://doi.org/10.1016/0895-4356(90)90060-3

12. Dayeh, T. A., Olsson, A. H., Volkov, P., Almgren, P., Rönn, T., & Ling, C. (2013). Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia*, *56*(5), 1036–1046. http://doi.org/10.1007/s00125-012-2815-7

13. dbGaP. (2012, March 12). Genome-Wide Association Study of Parkinson Disease: Genes and Environment dbGaP Study Accession: phs000196.v1.p1 . Retrieved from http://www.ncbi.nlm.nih.gov: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000196.v1.p1

14. Desplats, P., Spencer, B., Coffee, E., Patel, P., Michael, S., Patrick, C., … Masliah, E. (2011). -Synuclein Sequesters Dnmt1 from the Nucleus: A NOVEL MECHANISM FOR EPIGENETIC ALTERATIONS IN LEWY BODY DISEASES. *Journal of Biological Chemistry*, *286*(11), 9031–9037. http://doi.org/10.1074/jbc.C110.212589

15. Di, Y. P., Zhao, J., & Harper, R. (2012). Cigarette smoke induces MUC5AC protein expression through the activation of Sp1. *The Journal of Biological Chemistry*, *287*(33), 27948–27958. http://doi.org/10.1074/jbc.M111.334375

16. Disney, Julian, Baldry, Eileen, Calma, Tom, & Briskman, Linda. (2011, October). Occasional Paper: What is Social Justice. National Pro Bono Research Center. Retrieved from https://wic041u.server-secure.com/vs155205_secure/CMS/files_cms/Occ_1_What%20is%20Social%20Justice_FINAL.pdf

17. Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, *11*(6), 446–450. http://doi.org/10.1038/nrg2809

18. Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., … Relton, C. L. (2014). Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clinical Epigenetics*, *6*(1), 4. http://doi.org/10.1186/1868-7083-6-4

19. Eyre, H., Kahn, R., & Robertson, R. M. (2004). Preventing Cancer, Cardiovascular Disease, and Diabetes: A common agenda for the American Cancer Society, the American Diabetes Association, and the American Heart Association. *Diabetes Care*, *27*(7), 1812–1824. http://doi.org/10.2337/diacare.27.7.1812
20. Fleischacker, S. (2005). *A short history of distributive justice* (1. paperback ed). Cambridge, Mass.: Harvard Univ. Press.

21. Fraser, N. (2003). *Redistribution or recognition?: a political-philosophical exchange*. London ; New York: Verso.

22. Frohlich, K. L., & Potvin, L. (2008). Transcending the Known in Public Health Practice: The Inequality Paradox: The Population Approach and Vulnerable Populations. *American Journal of Public Health*, *98*(2), 216–221. http://doi.org/10.2105/AJPH.2007.114777

23. Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, *32*(4), 361–369. http://doi.org/10.1002/gepi.20310

24. Garte, S. (2006). Dose effects in gene environment interaction: an enzyme kinetics based approach. *Medical Hypotheses*, *67*(3), 488–492. http://doi.org/10.1016/j.mehy.2006.03.018

25. Gasser, T., Hardy, J., & Mizuno, Y. (2011). Milestones in PD genetics. *Movement Disorders*, *26*(6), 1042–1048. http://doi.org/10.1002/mds.23637

26. Ghazarian, A. A., Simonds, N. I., Bennett, K., Pimentel, C. B., Ellison, G. L., Gillanders, E. M., … Mechanic, L. E. (2013). A Review of NCI's Extramural Grant Portfolio: Identifying Opportunities for Future Research in Genes and Environment in Cancer. *Cancer Epidemiology Biomarkers & Prevention*, *22*(4), 501–507. http://doi.org/10.1158/1055-9965.EPI-13-0156

27. Gibb,W.L. (1988). The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry , 51:745-752.*

28. Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nature Reviews Genetics*, *9*(8), 575–581. http://doi.org/10.1038/nrg2383

29. Gorell, J. M., Johnson, C. C., Rybicki, B. A., Peterson, E. L., & Richardson, R. J. (1998). The risk of Parkinson's disease with exposure to pesticides, farming, well water, and rural living. *Neurology*, *50*(5), 1346–1350. http://doi.org/10.1212/WNL.50.5.1346

30. Hampton, J. (1997). *Political philosophy*. Boulder, Colo: Westview Press.

31. Hamza, T. H., Chen, H., Hill-Burns, E. M., Rhodes, S. L., Montimurro, J., Kay, D. M., … Payami, H. (2011). Genome-Wide Gene-Environment Study Identifies Glutamate Receptor Gene GRIN2A as a Parkinson's Disease Modifier Gene via Interaction with Coffee. *PLoS Genetics*, *7*(8), e1002237. http://doi.org/10.1371/journal.pgen.1002237

32. Han, L., Lin, I. G., & Hsieh, C.-L. (2001). Protein Binding Protects Sites on Stable Episomes and in the Chromosome from De Novo Methylation. *Molecular and Cellular Biology*, *21*(10), 3416–3424. http://doi.org/10.1128/MCB.21.10.3416-3424.2001

33. Handa, V., & Jeltsch, A. (2005). Profound Flanking Sequence Preference of Dnmt3a and Dnmt3b Mammalian DNA Methyltransferases Shape the Human Epigenome. *Journal of Molecular Biology*, *348*(5), 1103–1112. http://doi.org/10.1016/j.jmb.2005.02.044

34. Hardy, J., Cai, H., Cookson, M. R., Gwinn-Hardy, K., & Singleton, A. (2006). Genetics of Parkinson's disease and parkinsonism. *Annals of Neurology*, *60*(4), 389–398. http://doi.org/10.1002/ana.21022

35. Hernán, M. A., Takkouche, B., Caamaño-Isorna, F., & Gestal-Otero, J. J. (2002). A meta-analysis of coffee drinking, cigarette smoking, and the risk of Parkinson's disease. *Annals of Neurology*, *52*(3), 276–284. http://doi.org/10.1002/ana.10277

36. Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., … Esteller, M. (2013). DNA methylation contributes to natural human variation. *Genome Research*, *23*(9), 1363–1372. http://doi.org/10.1101/gr.154187.112

37. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367. http://doi.org/10.1073/pnas.0903103106

38. Holland, B. (2008). Justice and the Environment in Nussbaum's "Capabilities Approach": Why Sustainable Ecological Capacity Is a Meta-Capability. *Political Research Quarterly*, *61*(2), 319–332. http://doi.org/10.1177/1065912907306471

39. Huang, J., Okuka, M., Lu, W., Tsibris, J. C. M., McLean, M. P., Keefe, D. L., & Liu, L. (2013). Telomere shortening and DNA damage of embryonic stem cells induced by cigarette smoke. *Reproductive Toxicology*, *35*, 89–95. http://doi.org/10.1016/j.reprotox.2012.07.003

40. Hughes, A. J., Daniel, S. E., Ben-Shlomo, Y., & Lees, A. J. (2002). The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. *Brain*, *125*(4), 861–870. http://doi.org/10.1093/brain/awf080

41. Hulka, B. S., & Margolin, B. H. (1992). Methodological issues in epidemiologic studies using biologic markers. *American Journal of Epidemiology*, *135*(2), 200–209.

42. Hutter, Carolyn, & Mechanic, Leah. (2013). The Continued Importance of Research in Gene-Environment Interactions in 21st Century Epidemiology. NCI Epidemiology and

Genomics Division. Retrieved from http://blog-epi.grants.cancer.gov/2013/12/05/gene-environment-interactions/

43. Idaghdour, Y., & Awadalla, P. (2013). Exploiting Gene Expression Variation to Capture Gene-Environment Interactions for Disease. *Frontiers in Genetics*, *3*. http://doi.org/10.3389/fgene.2012.00228

44. Inoue, H., Lin, L., Lee, X., Shao, Z., Mendes, S., Snodgrass-Belt, P., … Isacson, O. (2007). Inhibition of the leucine-rich repeat protein LINGO-1 enhances survival, structure, and function of dopaminergic neurons in Parkinson's disease models. *Proceedings of the National Academy of Sciences*, *104*(36), 14430–14435. http://doi.org/10.1073/pnas.0700901104

45. Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*(4), 368–376. http://doi.org/10.1136/jnnp.2007.131045

46. Kadonaga, J. T., Carner, K. R., Masiarz, F. R., & Tjian, R. (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell*, *51*(6), 1079–1090.

47. Kaminsky, Z. A., Tang, T., Wang, S.-C., Ptak, C., Oh, G. H. T., Wong, A. H. C., … Petronis, A. (2009). DNA methylation profiles in monozygotic and dizygotic twins. *Nature Genetics*, *41*(2), 240–245.

48. http://doi.org/10.1038/ng.286

49. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. http://doi.org/10.1101/gr.229102. Article published online before print in May 2002

50. Kiyohara, C., & Kusuhara, S. (2011a). Cigarette smoking and Parkinson's disease: a meta-analysis. *Fukuoka Igaku Zasshi = Hukuoka Acta Medica*, *102*(8), 254–265.

51. Kiyohara, C., & Kusuhara, S. (2011b). Cigarette smoking and Parkinson's disease: a meta-analysis. *Fukuoka Igaku Zasshi = Hukuoka Acta Medica*, *102*(8), 254–265.
52. Klein, C., & Westenberger, A. (2012). Genetics of Parkinson's Disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(1), a008888–a008888. http://doi.org/10.1101/cshperspect.a008888

53. Korhonen, T., & Kaprio, J. (2011). Genetic Epidemiology of Smoking Behaviour and Nicotine Dependence. In John Wiley & Sons, Ltd (Ed.), *eLS*. Chichester, UK: John Wiley & Sons, Ltd. Retrieved from http://doi.wiley.com/10.1002/9780470015902.a0023476

54. Krieger, N. (1999). Questioning epidemiology: objectivity, advocacy, and socially responsible science. *American Journal of Public Health*, *89*(8), 1151–1153. http://doi.org/10.2105/AJPH.89.8.1151

55. Le Houezec, J. (2003). Role of nicotine pharmacokinetics in nicotine addiction and nicotine replacement therapy: a review. *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, *7*(9), 811–819.

56. Lee, E. W., & D'Alonzo, G. E. (1993). Cigarette smoking, nicotine addiction, and its pharmacologic treatment. *Archives of Internal Medicine*, *153*(1), 34–48.

57. Lee, K. W. K., & Pausova, Z. (2013). Cigarette smoking and DNA methylation. *Frontiers in Genetics*, *4*. http://doi.org/10.3389/fgene.2013.00132

58. Lesage, S., & Brice, A. (2009). Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Human Molecular Genetics*, *18*(R1), R48–R59. http://doi.org/10.1093/hmg/ddp012

59. Link, B. G., & Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior*, *Spec No*, 80–94.

60. Liu, C., Maity, A., Lin, X., Wright, R. O., & Christiani, D. C. (2012). Design and analysis issues in gene and environment studies. *Environmental Health*, *11*(1), 93. http://doi.org/10.1186/1476-069X-11-93

61. Liu, Q., Liu, L., Zhao, Y., Zhang, J., Wang, D., Chen, J., … Liu, Z. (2011). Hypoxia induces genomic DNA demethylation through the activation of HIF-1α and transcriptional upregulation of MAT2A in hepatoma cells. *Molecular Cancer Therapeutics*, *10*(6), 1113–1123. http://doi.org/10.1158/1535-7163.MCT-10-1010

62. Lix, L. M., Hobson, D. E., Azimaee, M., Leslie, W. D., Burchill, C., & Hobson, S. (2010). Socioeconomic variations in the prevalence and incidence of Parkinson's disease: a population-based analysis. *Journal of Epidemiology & Community Health*, *64*(4), 335–340. http://doi.org/10.1136/jech.2008.084954

63. M. V. Dogan. (2014, October). *Smoking Associated DNA Methylation Changes in Peripheral Blood Mononuclear Cells from African American Women and Weighted Protein-Protein Interaction Networks.* Poster presented at the ASHG, San Diego. Retrieved from http://www.ashg.org/2014meeting/abstracts/fulltext/f140120999.htm

64. Macintyre, G., Bailey, J., Haviv, I., & Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, *26*(18), i524–i530. http://doi.org/10.1093/bioinformatics/btq378

65. Maden, M. (2007). Retinoic acid in the development, regeneration and maintenance of the nervous system. *Nature Reviews Neuroscience*, *8*(10), 755–765. http://doi.org/10.1038/nrn2212

66. Maher, B. (2012). ENCODE: The human encyclopaedia. *Nature*, *489*(7414), 46–48. http://doi.org/10.1038/489046a

67. Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., … Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, *39*(10), 1181–1186. http://doi.org/10.1038/ng1007-1181

68. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. http://doi.org/10.1038/nature08494

69. Marklund, N., Bareyre, F. M., Royo, N. C., Thompson, H. J., Mir, A. K., Grady, M. S., … McIntosh, T. K. (2007). Cognitive outcome following brain injury and treatment with an

inhibitor of Nogo-A in association with an attenuated downregulation of hippocampal growth-associated protein-43 expression. *Journal of Neurosurgery*, *107*(4), 844–853. http://doi.org/10.3171/JNS-07/10/0844

70. Marmot, M. (2005). Social determinants of health inequalities. *The Lancet*, *365*(9464), 1099–1104. http://doi.org/10.1016/S0140-6736(05)71146-6

71. Masliah, E., Dumaop, W., Galasko, D., & Desplats, P. (2013). Distinctive patterns of DNA methylation associated with Parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics*, *8*(10), 1030–1038. http://doi.org/10.4161/epi.25865

72. Mayeux, R., Marder, K., Cote, L. J., Denaro, J., Hemenegildo, N., Mejia, H., … Gurland, B. (1995). The frequency of idiopathic Parkinson's disease by age, ethnic group, and sex in northern Manhattan, 1988-1993. *American Journal of Epidemiology*, *142*(8), 820–827.

73. Mazzoni, P., Shabbott, B., & Cortes, J. C. (2012). Motor Control Abnormalities in Parkinson's Disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(6), a009282–a009282. http://doi.org/10.1101/cshperspect.a009282

74. McCulloch, C. C., Kay, D. M., Factor, S. A., Samii, A., Nutt, J. G., Higgins, D. S., … Payami, H. (2008). Exploring gene-environment interactions in Parkinson's disease. *Human Genetics*, *123*(3), 257–265. http://doi.org/10.1007/s00439-008-0466-z

75. Mechanic, D. (2007). Population Health: Challenges for Science and Society. *The Milbank Quarterly*, *85*(3), 533–559. http://doi.org/10.1111/j.1468-0009.2007.00498.x

76. Mercer, B. A., Wallace, A. M., Brinckerhoff, C. E., & D'Armiento, J. M. (2009). Identification of a Cigarette Smoke–Responsive Region in the Distal MMP-1 Promoter. *American Journal of Respiratory Cell and Molecular Biology*, *40*(1), 4–12. http://doi.org/10.1165/rcmb.2007-0310OC

77. Miranda, T. B., & Jones, P. A. (2007). DNA methylation: The nuts and bolts of repression. *Journal of Cellular Physiology*, *213*(2), 384–390. http://doi.org/10.1002/jcp.21224

78. Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C., & Leonhardt, H. (2005). Recruitment of DNA methyltransferase I to DNA repair sites. *Proceedings of the National Academy of Sciences*, *102*(25), 8905–8909. http://doi.org/10.1073/pnas.0501034102

79. National Institutes of Health. (2014). *Mission*. NIH. Retrieved from http://www.nih.gov/about/mission.htm

80. Need, A. C., & Goldstein, D. B. (2009a). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, *25*(11), 489–494. http://doi.org/10.1016/j.tig.2009.09.012

81. Need, A. C., & Goldstein, D. B. (2009b). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, *25*(11), 489–494. http://doi.org/10.1016/j.tig.2009.09.012

82. Nielsen, S. S., Franklin, G. M., Longstreth, W. T., Swanson, P. D., & Checkoway, H. (2013). Nicotine from edible *Solanaceae* and risk of Parkinson disease: PD and Edible *Solanaceae*. *Annals of Neurology*, *74*(3), 472–477. http://doi.org/10.1002/ana.23884

83. Numata, S., Ye, T., Hyde, T. M., Guitart-Navarro, X., Tao, R., Wininger, M., … Lipska, B. K. (2012). DNA Methylation Signatures in Development and Aging of the Human Prefrontal Cortex. *The American Journal of Human Genetics*, *90*(2), 260–272. http://doi.org/10.1016/j.ajhg.2011.12.020

84. Nussbaum, M. C. (2011). *Creating capabilities: the human development approach*. Cambridge, Mass: Belknap Press of Harvard University Press.

85. Nussbaum, R. (1997). Genetics of Parkinson's disease. *Human Molecular Genetics*, *6*(10), 1687–1691. http://doi.org/10.1093/hmg/6.10.1687

86. Olson, K. R. (1984). Carbon monoxide poisoning: mechanisms, presentation, and controversies in management. *The Journal of Emergency Medicine*, *1*(3), 233–243.

87. Ottman, R. (1996). Gene-environment interaction: definitions and study designs. *Preventive Medicine*, *25*(6), 764–770.

88. Paltoo, D. N., Rodriguez, L. L., Feolo, M., Gillanders, E., Ramos, E. M., Rutter, J. L., … Green, E. D. (2014). Data use under the NIH GWAS Data Sharing Policy and future directions. *Nature Genetics*, *46*(9), 934–938. http://doi.org/10.1038/ng.3062

89. Parkinson's Disease Foundation. (2015). *Parkinson's Statistics*. Parkinson's Disease Foundation. Retrieved from http://www.pdf.org/en/parkinson_statistics

90. Pei-Chien Tsai. (2014, May). *Genome-wide Association Scans Identify Differentially Methylated and Expressed Regions Related to Smoking in Adipose Tissue*. Presented at the Illumina meeting, King's College, London. Retrieved from http://hermes.diskstation.me/~rob/450k/powerpoints/workshop3/3rd_450kmeeting_PCT.pdf

91. Pezzoli, G., & Cereda, E. (2013). Exposure to pesticides or solvents and risk of Parkinson disease. *Neurology*, *80*(22), 2035–2041. http://doi.org/10.1212/WNL.0b013e318294b3c8

92. Philibert, R. A., Sears, R. A., Powers, L. S., Nash, E., Bair, T., Gerke, A. K., … Monick, M. M. (2012). Coordinated DNA methylation and gene expression changes in smoker alveolar macrophages: specific effects on VEGF receptor 1 expression. *Journal of Leukocyte Biology*, *92*(3), 621–631. http://doi.org/10.1189/jlb.1211632

93. Powers, K. M., Kay, D. M., Factor, S. A., Zabetian, C. P., Higgins, D. S., Samii, A., … Payami, H. (2008). Combined effects of smoking, coffee, and NSAIDs on Parkinson's disease risk. *Movement Disorders: Official Journal of the Movement Disorder Society*, *23*(1), 88–95. http://doi.org/10.1002/mds.21782

94. Powers, M., & Faden, R. R. (2008). *Social justice: the moral foundations of public health and health policy*. Oxford: Oxford Univ. Press.

95. Priyadarshi, A. (2001). Environmental Risk Factors and Parkinson's Disease: A Metaanalysis. *Environmental Research*, *86*(2), 122–127. http://doi.org/10.1006/enrs.2001.4264

96. Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Research: The Official Journal of the International Society for Twin Studies*, *5*(6), 554–571. http://doi.org/10.1375/136905202762342026

97. Quik, M. (2004). Smoking, nicotine and Parkinson's disease. *Trends in Neurosciences*, *27*(9), 561–568. http://doi.org/10.1016/j.tins.2004.06.008

98. Quik, M., Parameswaran, N., McCallum, S. E., Bordia, T., Bao, S., McCormack, A., … Di Monte, D. A. (2006). Chronic oral nicotine treatment protects against striatal degeneration in MPTP-treated primates. *Journal of Neurochemistry*, *98*(6), 1866–1875. http://doi.org/10.1111/j.1471-4159.2006.04078.x

99. Quik, M., Perez, X. A., & Bordia, T. (2012). Nicotine as a potential neuroprotective agent for Parkinson's disease. *Movement Disorders*, *27*(8), 947–957. http://doi.org/10.1002/mds.25028

100. R Foundation for Statistical Computing. (2008). *R Development Core Team (2008). R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from http://www.R-project.org.

101. Rawls, John. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

102. Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., … Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, *411*(6834), 199–204. http://doi.org/10.1038/35075590

103. Ross, G. W., Petrovitch, H., Abbott, R. D., Nelson, J., Markesbery, W., Davis, D., … White, L. R. (2004). Parkinsonian signs and substantia nigra neuron density in decendents elders without PD. *Annals of Neurology*, *56*(4), 532–539. http://doi.org/10.1002/ana.20226

104. Rubin, M. S., Colen, C. G., & Link, B. G. (2010). Examination of Inequalities in HIV/AIDS Mortality in the United States From a Fundamental Cause Perspective. *American Journal of Public Health*, *100*(6), 1053–1059. http://doi.org/10.2105/AJPH.2009.170241

105. Rundle, A., & Schwartz, S. (2003). Issues in the epidemiological analysis and interpretation of intermediate biomarkers. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *12*(6), 491–496.

106. Samii, A., Etminan, M., Wiens, M. O., & Jafari, S. (2009). NSAID Use and the Risk of Parkinson's Disease: Systematic Review and Meta-Analysis of Observational Studies. *Drugs & Aging*, *26*(9), 769–779. http://doi.org/10.2165/11316780-000000000-00000

107. Satta, R., Maloku, E., Zhubi, A., Pibiri, F., Hajos, M., Costa, E., & Guidotti, A. (2008). Nicotine decreases DNA methyltransferase 1 expression and glutamic acid decarboxylase 67 promoter methylation in GABAergic interneurons. *Proceedings of the National Academy of Sciences*, *105*(42), 16356–16361. http://doi.org/10.1073/pnas.0808699105

108. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, *22*(9), 1748–1759. http://doi.org/10.1101/gr.136127.111

109. Scherf, D. B., Sarkisyan, N., Jacobsson, H., Claus, R., Bermejo, J. L., Peil, B., … Risch, A. (2013). Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRNB4. *Oncogene*, *32*(28), 3329–3338. http://doi.org/10.1038/onc.2012.344

110. Selamat, S. A., Chung, B. S., Girard, L., Zhang, W., Zhang, Y., Campan, M., … Laird-Offringa, I. A. (2012). Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Research*, *22*(7), 1197–1211. http://doi.org/10.1101/gr.132662.111

111. Sen, A. (1995). *Inequality Reexamined*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/0198289286.001.0001/acprof-9780198289289

112. Shen, J., & Yakel, J. L. (2009). Nicotinic acetylcholine receptor-mediated calcium signaling in the nervous system. *Acta Pharmacologica Sinica*, *30*(6), 673–680. http://doi.org/10.1038/aps.2009.64

113. Shigaki, H., Baba, Y., Watanabe, M., Iwagami, S., Miyake, K., Ishimoto, T., … Baba, H. (2012). LINE-1 Hypomethylation in Noncancerous Esophageal Mucosae is Associated with Smoking History. *Annals of Surgical Oncology*, *19*(13), 4238–4243. http://doi.org/10.1245/s10434-012-2488-y

114. Shoemaker, R., Deng, J., Wang, W., & Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, *20*(7), 883–889. http://doi.org/10.1101/gr.104695.109

115. Sidransky, E. (2006). Heterozygosity for a Mendelian disorder as a risk factor for complex disease. *Clinical Genetics*, *70*(4), 275–282. http://doi.org/10.1111/j.1399-0004.2006.00688.x

116. Singh, G. K., Azuine, R. E., & Siahpush, M. (2013). Widening Socioeconomic, Racial, and Geographic Disparities in HIV/AIDS Mortality in the United States, 1987–2011. *Advances in Preventive Medicine*, *2013*, 1–13. http://doi.org/10.1155/2013/657961

117. Smith, C. J., & Hansch, C. (2000). The relative toxicity of compounds in mainstream cigarette smoke condensate. *Food and Chemical Toxicology: An International Journal Published for the British Industrial Biological Research Association*, *38*(7), 637–646.

118. StataCorp. (2011). Stata Statistical Software: Release 12. StataCorp LP. College Station, TX.

119. Stewart, P. (1999). Challenges to retrospective exposure assessment. *Scandinavian Journal of Work, Environment & Health*, *25*(6), 505–510. http://doi.org/10.5271/sjweh.473

120. Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., … Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, *39*(10), 1217–1224. http://doi.org/10.1038/ng2142

121. Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics*, *187*(2), 367–383. http://doi.org/10.1534/genetics.110.120907

122. Suter, M., Ma, J., Harris, A. S., Patterson, L., Brown, K. A., Shope, C., … Aagaard-Tillery, K. M. (2011). Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics*, *6*(11), 1284–1294. http://doi.org/10.4161/epi.6.11.17819

123. Tanner, C. M., Ottman, R., Goldman, S. M., Ellenberg, J., Chan, P., Mayeux, R., & Langston, J. W. (1999a). Parkinson disease in twins: an etiologic study. *JAMA*, *281*(4), 341–346.

124. Tanner, C. M., Ottman, R., Goldman, S. M., Ellenberg, J., Chan, P., Mayeux, R., & Langston, J. W. (1999b). Parkinson disease in twins: an etiologic study. *JAMA*, *281*(4), 341–346.

125. Thomas, B., & Beal, M. F. (2007). Parkinson's disease. *Human Molecular Genetics*, *16*(R2), R183–R194. http://doi.org/10.1093/hmg/ddm159

126. Thorisson, G. A., Muilu, J., & Brookes, A. J. (2009). Genotype–phenotype databases: challenges and solutions for the post-genomic era. *Nature Reviews Genetics*, *10*(1), 9–18. http://doi.org/10.1038/nrg2483

127. Tofaris, G. K., & Spillantini, M. G. (2005). Alpha-synuclein dysfunction in Lewy body diseases. *Movement Disorders: Official Journal of the Movement Disorder Society*, *20 Suppl 12*, S37–44. http://doi.org/10.1002/mds.20538

128. Trinh, K., Andrews, L., Krause, J., Hanak, T., Lee, D., Gelb, M., & Pallanck, L. (2010). Decaffeinated Coffee and Nicotine-Free Tobacco Provide Neuroprotection in Drosophila Models of Parkinson's Disease through an NRF2-Dependent Mechanism. *Journal of Neuroscience*, *30*(16), 5525–5532.

http://doi.org/10.1523/JNEUROSCI.4777-09.2010

129. Turner, JH. (1997). *The institutional order: Economy, kinship, religion, polity, law, and education in evolutionary and comparative perspective.* New York: Longman. Retrieved from http://www.worldcat.org/title/institutional-order-economy-kinship-religion-polity-law-and-education-in-evolutionary-and-comparative-perspective/oclc/35138516

130. United Nations. (2013). *World Population Ageing*. United Nations, Department of Economic and Social Affairs Population Division. Retrieved from http://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2013.pdf

131. UNRISD. (2013, June). Health Systems as Social Institutions: Progress towards Health in All Policies. United Nations. Retrieved from http://www.unrisd.org/80256B3C005BB128/(httpProjects)/DB13A3F75C30A5B7C1257A1000576F5F?OpenDocument

132. Van Den Eeden, S. K. (2003a). Incidence of Parkinson's Disease: Variation by Age, Gender, and Race/Ethnicity. *American Journal of Epidemiology*, *157*(>11), 1015–1022. http://doi.org/10.1093/aje/kwg068

133. Van Den Eeden, S. K. (2003b). Incidence of Parkinson's Disease: Variation by Age, Gender, and Race/Ethnicity. *American Journal of Epidemiology*, *157*(>11), 1015–1022. http://doi.org/10.1093/aje/kwg068

134. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics*, *90*(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

135. Wang, K., Dickson, S. P., Stolle, C. A., Krantz, I. D., Goldstein, D. B., & Hakonarson, H. (2010). Interpretation of Association Signals and Identification of Causal Variants from Genome-wide Association Studies. *The American Journal of Human Genetics*, *86*(5), 730–742. http://doi.org/10.1016/j.ajhg.2010.04.003

136. Weeks, D. E., & Lathrop, G. M. (1995). Polygenic disease: methods for mapping complex disease traits. *Trends in Genetics: TIG*, *11*(12), 513–519.

137. Wright, R. O., & Christiani, D. (2010). Gene–environment interaction and children's health and development: *Current Opinion in Pediatrics*, *22*(2), 197–201. http://doi.org/10.1097/MOP.0b013e328336ebf9

138. Wright Willis, A., Evanoff, B. A., Lian, M., Criswell, S. R., & Racette, B. A. (2010). Geographic and Ethnic Variation in Parkinson Disease: A Population-Based Study of US Medicare Beneficiaries. *Neuroepidemiology*, *34*(3), 143–151. http://doi.org/10.1159/000275491

139. Yamada-Fowler, N., Fredrikson, M., & Söderkvist, P. (2014). Caffeine Interaction with Glutamate Receptor Gene GRIN2A: Parkinson's Disease in Swedish Population. *PLoS ONE*, *9*(6), e99294. http://doi.org/10.1371/journal.pone.0099294

140. Yavich, L., Jäkälä, P., & Tanila, H. (2006). Abnormal compartmentalization of norepinephrine in mouse dentate gyrus in ?-synuclein knockout and A30P transgenic mice. *Journal of Neurochemistry*, *99*(3), 724–732. http://doi.org/10.1111/j.1471-4159.2006.04098.x

141. Young, I. M. (2013). *Responsibility for justice* (First issued as an Oxford Univ. Press paperback). Oxford: Oxford Univ. Press.

142. Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., … Illig, T. (2013). Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLoS ONE*, *8*(5), e63812. http://doi.org/10.1371/journal.pone.0063812

143. Zhang, B., Zhu, W., Yang, P., Liu, T., Jiang, M., He, Z.-N., … Chen, W. (2011). Cigarette Smoking and p16INK4α Gene Promoter Hypermethylation in Non-Small Cell Lung Carcinoma Patients: A Meta-Analysis. *PLoS ONE*, *6*(12), e28882. http://doi.org/10.1371/journal.pone.0028882