# Table of Contents

1. Introduction .................................................................................................................. 2
2. Sociogenomic analysis of life attainments .................................................................. 3
3. Large-scale genomic meta-analysis identifies loci harbouring genes for human reproductive behavior .................................................................................................................. 7
4. What data do we need to study genes and complex human phenotypes? ............ 10
5. Implications of the research for biomedical science .............................................. 13
6. Mortality selection in a genetic sample and implications for association studies .... 16
7. GCTA and all that .......................................................................................................... 19
8. Random-effects estimates of heritability for simple and complex traits: some statistical theory ................................................................................................................................. 23
9. Statistical learning approaches to identifying gene-environment interaction ........ 25
11. Do PCs overcontrol causal effects ............................................................................... 31
12. Genetic load and polygenic scores ............................................................................... 34
13. Genetic heterogeneity in human fertility and educational attainment by country and birth cohort ........................................................................................................................................... 37
15. Implications of the research for social science ......................................................... 43
16. Development of a measure of cognitive age .............................................................. 46
17. Discussion: The aging process – how do we measure it in this setting? How does it influence results derived from genetic studies? .................................................................................. 52
18. Closing Remarks .......................................................................................................... 54

Appendix A: Program ......................................................................................................... 55
Appendix B: Meeting Participants ..................................................................................... 58
1. Introduction

Opening remarks, Shripad Tuljapurkar

In general, we observe a phenotype of an individual at some age and time, and this is a function of several things:

1. The phenotypic history of the individual (i.e. other phenotypes, or the same phenotype of an individual at a younger age)
2. The external environment (individual in ontogeny, family, society)
3. The internal environment (microbiome)
4. The genotype (expression, epigenetic state, network state)

This entire process integrates over history in some way.

The phenotypic history and the individual in ontogeny, I think you all are familiar with. For the external environment, there are hierarchies. We have the individual environment, family, society, and so forth. For the genotype, we tend to think of genes as handed down, but genes are actually dynamic entities that change expression through the course of life; they change epigenetic state, and finally – and this is largely ignored – they change network state. That is, we are born with a set of genes, but these genes do not have the same function across a lifespan. Early in life, certain genes are involved in development, but later in life these genes have different functions.

Typically, we think of phenotype as something we can measure (e.g. blood pressure, height, weight). Then, we have some outcome of interest (e.g. fertility) and we want to see what relationship might exist between the phenotype and outcome of interest. It has become very common to use Genome Wide Association Studies (GWAS) to do this kind of analysis. In these studies, single nucleotide polymorphisms (SNPs) are examined. A number of SNPs are chosen in a number of people for a number of phenotypes. The number of SNPs chosen is technology driven; at the present time, we can choose from about a few million to a billion SNPs, though the genome has well over a billion SNPs. The number of people chosen is limited by cost, as studies are grant driven. Finally, the number of phenotypes chosen are limited by sample design, cost, and permissions.

There is a lot of attention given to heritability, which is perhaps overblown. Heritability describes how much of the observed phenotypic variation is due to genes in some specific context, so it is actually an estimator under fairly specific conditions. As a result, heritability is not usually a useful or reliable guide to evolution. Furthermore, estimation is problematic, and there is no or little biology involved in the process.

The problem is that despite high heritability, high genetic variation, and strong evidence of selection, there are very few cases of micro-evolutionary change. This suggests that we may be looking at the wrong thing, or there is selection of correlated things that haven’t been probed, or something of the sort, and it’s a puzzle. However, according to a notable study, “human GWAS studies might lead to richer insights if they incorporated the molecular architecture of complex traits” (Remington 2015).

In terms of the molecular architecture, where are these SNPs, and how are these loci biologically relevant? This can be done by taking a large number of phenotypes, finding which
SNPs are significant and identifying their location, and looking at chromosomal regions where genes may be linked, and see what these genes do. It is also important to consider if these genes are regulatory, that is, if they regulate other genes, or if they are structural genes that act as catalysts or shunts. According to one study by Hunter Fraser, most SNPs that affect genes are in regulatory regions. That is an interesting phenomenon because you can use these locations to formulate testable hypotheses about 1) gene function 2) genetic correlations between phenotypes 3) environmental correlation between phenotypes.

There is a lot of work still to be done in the field. One direction to go is in comparing different countries. If you look at populations within countries, a lot remains unclear due to diversity (or lack thereof) of that population, and what role migration may play in introducing variation to the gene pool. Another area of interest is GxE interactions. By doing comparative work, we can start probing these interactions. The trouble with the current work is that we do not have good measures of the environment. The Add Health data provides potentially new and strong models for environment, as it follows individuals throughout their life. The question still remains, however, if there is comparable data for other countries.

There are many questions related to the following issues on models, which have not received much attention:

1. Dispersal
2. GRS and the additive model (one trait versus many, recombination)
3. Non-additive GRS
4. GxE models

Finally, questions remain on the topic of relatedness:

1. Time scales
2. Identity by descent versus time scale
3. Pedigrees, base pedigree, controls
4. What do PC “corrections” mean

These are questions I hope we can discuss during the course of this meeting.

2. Sociogenomic analysis of life attainments

Daniel Belsky

Introduction

My research concerns the causes of socioeconomic gradients in health and the aging process. I study genetics because the features of human development that influence disease and disability processes are heritable; they’re influenced by genetic differences between individuals. A second reason I study genetics is that DNA is special; it comes before almost everything else humans experience. The idea behind this project is that by understanding how DNA regulates human behavior and achievement, we can learn something about causal processes that can
provide targets for interventions that will apply broadly across the population, regardless of genotype.

I want to start with some background on how I think about genetics. Until recently, research to uncover molecular genetic causes of human phenotypes was hypothesis driven. The method was to reverse-engineer the biology of a phenotype to identify a protein or molecule of interest, use that protein/molecule to identify a gene, and then screen that gene for variation that could then be tested for association with the phenotype. We called that the “candidate gene” approach. These days, with the aid of new technology, we can take a comprehensive survey of the genome and do hypothesis free search to identify variants linked with phenotypes. The variants we study most often in this hypothesis-free work are single-letter changes in the DNA sequence that occur in at least 1% of the population, called single nucleotide polymorphisms (SNPs). This hypothesis free search for SNPs, called a “genome-wide association study” (GWAS) is the basis for the work I will present today.

GWAS is a brute force approach to discovery. Millions of SNPs are tested for association with a phenotype of interest. The resulting correction for multiple testing is severe, placing high demand on statistical power in GWAS design. Theory and empirical evidence suggest the best way to increase power in GWAS is to accumulate very large samples, even at the expense of having relatively coarse phenotypes. In the social sciences, this means studying phenotypes like educational attainment, which can be measured easily in 100s of thousands of individuals, instead of psychological characteristics or economic behaviors that might be more precise quantifications of the phenotype of interest. The need for extremely large samples also precludes analysis of detailed information about those samples – where and when they grew up, how they lived, and so on. This leaves many unanswered questions about mechanisms connecting genotype and phenotype.

Learning from GWAS by “working from the top down.” The traditional approach to following up GWAS discoveries works from the bottom up, tracing a molecular path from DNA sequence, through RNA transcription, and on up to phenotype. The goal of bottom-up work is usually to find a molecule that can serve as a drug target. In my work, I start from the top and work down. The goal is to trace a path backward from the GWAS phenotype to its developmental antecedents. The goal of this top-down research is to find a behavior of developmental process that can serve as a target for a non-drug intervention like a policy or a program.

How do we do this top-down work? Individual genetic variants have only very small effects, but many thousands of variants contribute to the sorts of human variation we are interested in studying. So to conduct effective follow-up, methods are needed to summarize genetic influences from across the genome. The method we use is called polygenic scoring. Polygenic scoring begins by using results from GWAS to assign weighted values to large numbers of variants in a person’s genome. Then, those weighted values are summed to form that person’s polygenic score. The result is a normally distributed measure that quantifies the level of genetic load toward a particular trait or outcome. All this background is discussed in greater detail in Belsky and Israel 2014, Biodemography & Social Biology.

In previous work, we and others have established that a polygenic score based on GWAS of educational attainment (1) predicts educational attainment in many different samples spanning different historical periods and geographic/socio-political contexts; and (2) is more than an artifact of socially-advantaged ancestry. This second point is important. Children inherit genes and social position from their parents. It could have been the case that the genetics of educational
attainment were simply the genetic signature of a privileged social class. They are not. We know this because when we compare siblings in the same family – people who share identical ancestry, we find that the sibling with the higher polygenic score tends to complete more years of schooling. So the polygenic score for educational attainment is measuring a real, substantive genetic influence. Today, I will present some work on what we think that influence might be.

Before going forward, I want to acknowledge that the genetics of socioeconomic attainment are a volatile topic and research is easily misinterpreted or misconstrued. To be clear, our work aims to uncover features of human development and behavior that contribute to socioeconomic success, with the aim of improving outcomes for everyone in the population, regardless of their genotype.

Study design

We studied life-course development of socioeconomic attainment in a cohort of N=918 individuals followed from birth in 1972-3 through their 38th birthday with 95% retention: the Dunedin Study. We measured life-course development using a combination of direct observations, psychometric testing, and interviews with Study members and their parents, informant surveys, and electronic record searches of government and credit-bureau databases.

We conducted polygenic scoring based on published GWAS results for educational attainment from the Social Science Genetic Association Consortium. Our aim was to uncover how genetic influences might shape educational trajectories and to what extent they might influence outcomes beyond education.

Results

We calculated polygenic scores for Dunedin Study members from about 2.3m SNPs present in our genetic database and the GWAS results published by SSGAC. We adjusted all analyses for the first ten principal components estimated from the genome-wide SNP data to account for potential population stratification.

The education polygenic score was normally distributed in the Dunedin cohort. Replicating previous studies, Dunedin Study members with higher polygenic scores tended to have completed more schooling \((r = 0.15, p<0.001)\).

We next examined Study members’ socioeconomic attainments through midlife. We compiled dossiers of attainment based on electronic record searches (social welfare benefit use, credit scores), interviews (income, assets, financial problems), and information about their occupations. Attainments were correlated. We computed an attainment factor score to test genetic influence on socioeconomic success through midlife. Children with higher polygenic scores grew up to achieve more socioeconomic success \((r = 0.13, p<0.001)\). About half of this genetic effect was explained by differences in education.

Because children inherit both genes and social position from their parents, we next tested whether genetic associations with socioeconomic attainment could be explained by children’s social origins. We quantified childhood social class from information about parents’ occupations collected at repeated waves when Study members were children. Children with higher polygenic scores tended to come from wealthier families. Children with lower polygenic scores tended to come from poorer families. Despite this gene-environment correlation, children with higher polygenic scores tended to achieve more socioeconomic success by midlife regardless of their social background \((r=0.11, p<0.001)\). Specifically, children with higher polygenic scores were more likely to be upwardly socially mobile, to achieve more, even when they were born poor.
Further analysis examined developmental and behavioral mechanisms underpinning genetic influence on social mobility. Beginning early in life, children with higher polygenic scores started to distinguish themselves from peers. They spoke at earlier ages, learned to read sooner, and went on to do better in school. After school, these children more often sought experiences overseas and, in fact, were more likely to emigrate (defined as residing outside of NZ for at least the past year at the time of most recent data collection when they were aged 38 years). Children with higher polygenic scores tended to select life partners with more education and higher income. And by midlife they were better at taking care of their own finances.

Finally, we examined psychological characteristics that might mediate the pattern of life course success we observed for children with higher polygenic scores. As expected, children with higher polygenic scores had higher IQ scores. But they also distinguished themselves in other domains. They had better self-control and were more interpersonally skillful (cooperative, friendly, communicative, etc.). Collectively, these cognitive and non-cognitive characteristics accounted for around half of genetic associations with life course attainments. Strikingly, children’s polygenic scores were not related to their physical health – as measured from blood pressure, lung function, height, weight, balance, and reviews of medical dossiers.

Conclusions

GWAS discoveries for education are not about education only. Instead, genetic discoveries for educational attainment correspond to a pattern of characteristics that manifest beginning early in life as accelerated language acquisition and mental development, mature into
academic achievement in school, and extend through patterns of career formation and mate selection, yielding a life course pattern of upward socioeconomic mobility socioeconomic success by midlife. In addition to intelligence, genetic influences on life course socioeconomic success are mediated by better self-control and interpersonal skill, but not better childhood physical health.

Next steps include tests of replication of findings in different contexts – other cohorts, other countries, other policy regimes – and investigation of environmental factors that may amplify or mitigate genetic influences.

Effect sizes were small. Use of genetic testing to engineer “precision” education is not yet possible. However, public conversation is needed about how genetics that correlate with social outcomes may be used in biomedicine and education.

Ultimately, the promise of work such as this lies in the identification of developmental and behavioral mechanisms that mediate genetic influence on socioeconomic success. These mechanisms can then serve as targets for interventions that can be delivered to anyone, regardless of genotype. For example, results from our study are consistent with the notion that intervention to accelerate language development and acquisition of reading skill may provide one path to promote upward social mobility.


---

3. Large-scale genomic meta-analysis identifies loci harbouring genes for human reproductive behavior

**Melinda Mills**

**Introduction**

A GWAS of human reproductive behavior was conducted to examine genetic variants of reproduction. Reproductive behavior was operationalized using the following phenotypes: age at first birth (AFB) and number of children ever born (NEB). In other words, we tried to answer the following question: is fertility in our genes? Our research has implications for demography, sociology, medical science, and evolutionary biology.

**Background**

Previous research, including many twin studies, have focused on reproductive outcomes, yet few genetic loci for reproductive behavior have been identified, and the biological mechanisms mediating said behavior are poorly understood. Little is also known about the

---

1 The material presented here was subsequently published in *Nature Genetics*. The link to the article is: [http://dx.doi.org/10.1038/ng.3698](http://dx.doi.org/10.1038/ng.3698). Frequently Asked Questions about the paper for press and the public is available at [http://www.sociogenome.com/data/NG2016FAQ/](http://www.sociogenome.com/data/NG2016FAQ/). A film from Oxford explaining the research is at [https://www.youtube.com/watch?v=PWSfWSb5KwE](https://www.youtube.com/watch?v=PWSfWSb5KwE)

**GENES AND COMPLEX HUMAN PHENOTYPES | 7**
genetic variants related to reproductive behavior in males, as data in this population is difficult to obtain.

Couples have been postponing their AFB and NEB, particularly in Europe. Later AFB may contribute to infertility in couples, partially due to increased ovulation defects, spermatogenic failure, and single or polygenic defects. The genetic causes of infertility are unknown, but are important to understand as infertile couples account for 10-15% of all couples. Genetics are not the only thing that influence human reproduction. Biological fecundity (i.e. length of reproductive period, ovulation), reproductive choice (i.e. personality, latent individual and partner characteristics), and environment also influence reproductive outcomes. These factors, in contrast to genetic factors, have been explored extensively by the scientific community. However, twin studies have indicated that 25-40% of reproductive behavior is heritable, and that 15% of the observed variance in AFB and 10% of the observed variance in NEB is explained by genetic variation.

Population and Study Design

Sixty three datasets from the U.S., Canada, Northern, and Western Europe were used, with data retrieved from various sources. The total sample size for AFB was around 238,000 (190,000 women, 48,000 men), and was limited to those individuals who had previously had a child. The total sample size for NEB was around 329,000 (225,000 women and 104,000 men). The large sample makes this study a well-powered one.

A GWAS on human reproductive behavior as operationalized by AFB and NEB was performed to identify genetic markers of reproductive behavior (SNPs). A multiple linear regression (additive model) was then performed to examine the combined effects of genetic variants on reproductive behavior for the population as a whole, and for each sex. Additional methods performed included: multi-trait analysis, population stratification, sex-specific genetic effects, polygenic score prediction, genetic correlations with other traits, biological follow-up, and a bivariate Heckman selection model.

Results

Twelve independent loci were significantly associated with NEB or AFB, 10 of which were novel loci. From the biological follow-up studies, 4 additional novel loci were discovered (3 AFB, 1 NEB). These loci harbor genes that play a role in gene expression in reproduction and fertility. The figures below show Manhattan plots of SNPs for AFB (left) and NEB (right).
We then tried to understand the biological mechanisms behind these genetic loci (e.g., are these genes involved in gene methylation, do they have regulatory consequences, are they in linkage disequilibrium with other genes, do they influence gene expression via pathways, etc)? We discovered that these loci influence gene expression via several pathways. Using multiple techniques, 50 SNPs were determined to be functional. 2 sets stand out: Chr 1 (18) and Chr 3 (25 SNPs) seem to play roles in active regulation and transcription of proteins related to sperm differentiation and ova/ovum.

Then, a literature-based search was performed on the 12 top hits from the GWAS to see if any studies have been conducted connecting this genes in a causal manner to reproduction and fertility. For men, there were hits related to spermatid differentiation (CREB3L4), spermatid maturation and acrosome reaction (HYAL3), and spermatogenesis (RBM5; CYHR1; GPT; RECQL4; PPP1R16A). Hits were also associated with:

- Fertility in female mice (EFNA5)
- Hormones related to fertility (HCN1)
- Reproductive tissues such as the ovum, oocyte, fallopian tube, and prostate (MST1R; CRTC2)
- Estrogen responsive gene, which aids in sexual maturation and development, as well as male fertility estrogen transfer (ESR1)

We next examined gene prioritization to see if there were specific genes that would cause these sorts of association? We found eight genes that were prioritized for AFB and NEB. Of these genes, 3 genes were likely to be causal for AFB (MON1A, RBM6, U73166.2). Moving to polygenic score prediction, the mean predictive power was 0.09% for AFB and 0.02% for NEB. These effects are small, but significant. For example, 1 SD increase in NEB polygenic score was associated with a 9% decrease in probability for women to remain childless. Similarly, 1 SD increase in the PGS for AFB was associated with an 8% reduction in the NEB hazard ratio in women. Finally, a 1 SD increase in AFB PGS was associated with a 3% decrease in age at natural menopause and 20 day increase in age at menarche. Bivariate LD score regression was then used to estimate pairwise genetic correlations with 27 publicly-available traits. There was a strong overlap between AFB and education as well as AFB and age at first intercourse.

Conclusion

To our knowledge, this is the largest GWAS on human reproduction to date. 12 genetic loci (10 novel) were identified as related to AFB or NEB. Chromosome 1 and Chromosome 3 were of particular interest. SNPs in LD in Chr 1 likely are involved with functional genes that encode 1) CREB, which at the protein level acts as a critical signal mediator for follicle-stimulating hormone (FSH) and 2) CREB3L4, which is expressed in the prostate, ovaries, uterus,
and testis and plays a role in spermatid differentiation. SNPs in LD in Chr 3 likely are involved with functional genes associated with methylation status and expression of CRTC2, which is implicated in polycystic ovarian syndrome. Overall, while the predictive power of the GWAS is quite low – the PGS score is a fraction of what is found in twin and family studies – the findings are still significant. Non-additive genetic effects, epistatic effects, rare variants and inflated estimates from twin studies due to shared environment, could all contribute to the low predictive power observed.

Questions/Discussion

Shripad Tuljapurkar suggests that a future direction for Mills’ research could be at looking breast cancer risk. If you look at breast cancer, many genes implicated in breast cancer are not BRCA-related. Looking at AFB, age at menarche, and NEB can provide interesting insight on the etiology of breast cancer. Melinda Mills notes that the group actually joined forces with an endometriosis center to see if there are any overlap from this GWAS with breast cancer.

Ken Wachter asks about control variables, given that there are a very limited number of control variables that are assessed across the spectrum of GWAS studies. Melinda Mills acknowledges that controls are a critical consideration, and that her other study adjusts for birth cohort and country to account for some socially mediated factors that may influence AFB and NEB.

Sid Kumar notes if population have been split to see for differences in allocation according to birth order, sex.

4. What data do we need to study genes and complex human phenotypes?

Kathleen Mullan Harris

Introduction

Today we will be talking about the evolution of genetic idea in Add Health and the research it produced, challenges in producing genetic data, what geneticists and social scientists care about, and finally current limitations in data collection.

Background
The scientific objective of Add Health was to understand the causes of health and health behavior in adolescents, with a focus on social context. When Add Health first began in the 1990s, it focused on the individual. Samples were drawn from individual schools, which allowed for characterization of the school environment and the ability to interview individuals within the sample that knew one another, including romantic partners, family, and friends. Data was collected in waves during in-home interviews, as illustrated on the right.

Data

From the outset of Add Health, it was important to include a biological aspect, but the main focus at the time was trying to understand environmental effects on health. As Add Health progressed, genetic data evolved with it. During Wave I, there was an embedded sample of 300 pairs of adolescents with some biological resemblance. This allowed for some basic behavioral genetics models, and produced one of the first GxE interactions, as shown below:

![Graph showing the relationship between genetic and shared environmental variance components by level of parental education.](image-url)

*Figure 1* Relations between genetic and shared environmental variance components by level of parental education. Standard errors at the means and the extremes are indicated by vertical lines. Arrows point to the estimates of heritability and shared environmental influences at the sample mean.
Here, one observes that as parental education increases, genetics play an increasing role on vocabulary compared to the environment. At the low end of SES, the environment is more important, suggesting that there is a role for intervention.

During Wave II in 1996, Add Health became the first socio-demographic study to collect DNA in the US. DNA collection kits were made to check for zygosity on 89 twin pairs. Then during Wave III in 2001, molecular data were collected on twins and full siblings (N~2600) using DNA kits that used the mouthwash method to retrieve buccal cell DNA. Such data allowed for the identification of candidate genes in dopamine and serotonin pathways, and generated some of the first genetic analysis with molecular data.

We faced many challenges in collecting and making use of all of this data. First, it was important to keep the theoretical motivation in mind when collecting genetic data (i.e. understanding the biological mechanisms behind genes), which can easily get lost during study design. We also faced challenges in data collection methods. Sometimes, an interview could not be conducted at home, so innovations needed to be made, including interviewing at other locations. Furthermore, the logistics of conducting a national field study are difficult. Hours are spent on logistics. For example, we collected urine to test for STIs, but this urine had a 48-hour window in which the sample must be returned, making samples from Alaska and Hawaii virtually unusable. In addition, the idea that lab assay results could be linked to 20 years of data on an individual faced resistance. Cost was another issue. During Wave IV in 2008, DNA was collected from participants’ saliva using oragene kits. Logistically, we had to take every one of these samples and record them. When the sample is big, the cost is big. DNA collection in Wave IV and genotyping cost around $2.5 million. Finally, the quality of research was variable. Though there are thoughtful, collaborative, and interdisciplinary approaches that have produced provocative findings, other studies saw social scientists conducting research with no genetic training and vice versa.

So why do we care about this data?
1) Environments and social experiences: the ways environment may moderate genetic effects on phenotype(s) and vice versa (GxE interactions)
2) Mechanisms: what biological pathways link social exposure to phenotype
3) Causal effects: the extent of the endogeneity of the environment, and the policy implications

However, I worry that there is not enough cross-talk between geneticists and social and behavioral scientists. I am worried that we are losing sight of where geneticists are going. Cross-training between social scientists and geneticists is of great importance.

Finally, I wanted to discuss the current limitations we face in our research. Most GWAS are based on all-white samples – we need more diversity in our GWAS studies. Researchers also often employ an a-theoretical approach, but if you torture your data enough, it will speak. In addition, there may be too much of a focus on discovery and having results. Finally, there is the thought that a large sample size will solve all of our problems, but having a larger sample size will not fix a poorly designed study.

Conclusion

Most GWAS are based on specific populations within a specific age group. However, it is important to study age and race variations. Relationships between variables, for example, have got to vary by age. We also need to develop better measures of complex phenotypes, characterize biological pathways rather than focus on the discovery of genetic variants, and return to a scientific framework in approaching these studies. From a social science perspective,
the most compelling directions for future research are social genomics and epigenetics. These two fields emphasize social exposures, and it is important to get at the mechanisms that mediate the relationship between genetics and environment. In terms of Add Health, the future of the dataset includes transcriptome data, which was just funded (N=5,000), and methylation data (N=5,000), which a grant is currently in progress.

Questions/Discussion

Sid Kumar asks how related the biological adolescent pairs actually are and what filter was used to select schools? Kathleen Mullan Harris clarifies that these pairs are some sort of sibling, including MZ and DZ twins. 80 high schools were selected from across the nation and matching feeder middle schools. We went in and interviewed everyone in that school. We sampled 20,000 participants from the total population of 100,000 students.

Dalton Conley remarks that Add Health is the one of the best national U.S. representative sample. A dataset like HRS, for example, has some limitations because one needs to survive to a certain age to be included. The Add Health dataset could contribute to our understanding of the deep history of reproduction. Many critics ask: why are we interested in genetics as social scientists. One of the best ways to defend this is to look at genetics through social history. But I could not find much literature in the US on this topic. Is this because Add Health is one of the only resources to do this? Are you aware of sociologists doing this type of research? Kathleen Mullan Harris responds by saying that the only other young dataset is Cardia, which is not representative as it is only in 3 cities. I don’t know of others that are trying to do this type of research, and do not here people requesting this data for that purpose.

Melinda Mills comments that there is a lot of medical data, but not much social science data. She noted that on the Data Access committee in the UK where she sits, they restricted access to studies that examine homosexuality, intelligence, and criminality. They didn’t want the data to be used for that reason. Harris responds that they give the link, and they have access to all the data. It costs us more to customize in any way. Well maybe we would take out the criminal data, but it’s difficult taking.

Kumar notes that the ugly data is linking gene associations as causal, rather than associated; asks if this is common. Harris responds by saying this is common in bad science, and Belksy notes that this is bad science journalism.

5. Implications of the research for biomedical science

Mark Cullen

There has been a demonization of biomedical science, but I first wanted to give some perspective for population genetics from the biomedical science community. Second, I wanted to discuss the data explosion that is occurring now. In general, we are in better shape than we think are, and we are heading in an interesting direction that will allow for some exciting work in the future. The impact of linked datasets for the social scientist will be tremendous. Furthermore, social scientists are beginning to acquire new tools that will better allow for causal inference.

Yesterday, there was a big data conference at Stanford, and there was a meeting with Claudia Williams, the White House Liaison for the precision medicine initiative. The initiative will enroll a million volunteers, who will have extensive biologic sampling done, with electronic medical records linked. I asked about the extent of environmental and social science data in the initiative, and how that data is going to be collected. Claudia looked at me and said, “we need to
walk before we run”. And by that she meant we need to figure out the biology first. Running was trying to think about what we append to the data (i.e. social science data).

The conference underscored a progression of three ways biomedical people are thinking about genetic data:

1) People who are into precision medicine. These individuals identify variances that are actionable (drug, therapy/treatment), but are usually very expensive. For example, the notion that by classifying the genetics of tumors, we can target and treat them. This group took over the conversation in two minutes.

2) People who are into precision health. These individuals use the fantasy of 23andMe and the predictive notion of genetics. 23andMe commercializes genetics, by selling kits that people can use to assess their risk for various genetic disorders. There is direct consumer marketing, but very few physicians involved. The idea here is to forget the value of an expensive therapy, and instead emphasize the need to know about personalized biology behind long-term risks. Then a personalized recommendation based on their predicted long-term trajectory, can be carved out. The margins here, however, are quite small. Environment here is something that needs to be known, not because the environment or behavior should be change, but because it shows how strong or weak genes are as predictors.

3) People who accept neither viewpoint. These individuals acknowledge that these very tools (i.e. expensive treatments) have phenomenal opportunity if we have the right datasets and think about them in the right way. Only then can we really begin to see the life course. For example, how are social factors at every point modifying our original endowment from conception? The traditional notion that the brain is a homeostatic process is replaced by this new notion that the brain is an allostatic process. As the brain is bombarded with responses, changes that occur in the short-term can be damaging long-term (e.g. the brain may be mediating diseases at a cellular level through shortening telomeres).

Now, for the first time we have a glimmering of what is going on from the cellular level where diseases are actually occurring to exogenous factors and behavior. Social science is still a little bit of a mystery, but we now have at least some pathway by which the genetic endowment may be modified. This is an exciting time because we can actually measure a lot of exogenous factors such as behavior, physical affect, and the environment, which we have previously been unable to measure. We can examine each stage of life, and see how our genetic endowment is affected and vice-versa. Looking back even 10 years ago, this is way different from the paradigm of precision medicine. It is still important to know the molecular mechanisms behind these relationships because if we are going to think about social forces that are really going to make a difference, we need to know where they interact and what the impact of that is going to be. We are actually aligned with the biomedical community with regard to this research, we just have a different takeaway.

When thinking about the main studies and datasets that we have available, every study omitted some key area. The HRS, for instance, omitted physical health. Now, we can measure all these things (e.g. physical health may be measured using a FitBit), and we are just a few years away from measuring everyday people’s physical metrics, such as food consumption. Sociological factors are obviously harder to measure, but the rest of the data can now be measured relatively easily. In addition, we have incredible troves of information – most people’s medical records are now digitized. You would be surprised how quickly a hospital can get data
from you during an ER visit. All this data can be linked geospatially or to administrative data. Of course, there is a huge privacy issue, which may make causal inference difficult, but the fact is, there is so much data out there: social security, IRS, Census, etc.

With all this data, we can nail down childhood circumstances, which is pretty amazing. This is frankly the direction data are going. Yes, there are limitations, but if we are propelling in our efforts to link data, then we are in for better times. We should be the first ones in to mine this data religiously to learn about the life course.

Ben Seligman elaborates on the idea of linking data, suggesting that we should reach out to the biomedical community to link up data. Because of the Affordable Care Act, pretty much all hospitals have to move to electronic medical records, and there is also this subtle pushing towards the formation of ACOs. This creates competition between hospital systems to get market share and to follow patients through their life course. Hospitals are starting to track patients across the life course from beginning to end. Doctors are so used to giving drugs, but now that we follow patients outside of the hospital, environmental interventions are becoming more of a viable and important option. If we’re working at academic hospitals with medical data, we can link these sorts of data.

Sid Kumar remarks that linking data comes with a host of privacy issues. Daniel Belsky asserts that there is a distinction between using data for commercial purposes, and using it for research purposes. IRB protocols make research data highly regulated. Sid Kumar counters by noting that fraud happens when it comes to data access. Mark Cullen then poses the following question: what if we had access to Safeway purchasing data of consumers? While there are a lot of anxieties about this data being collected, having this data also solves many gaps in constructing social history. Kathleen Harris remarks that there are IRB issues with linking data. IRS data would give us the address of a respondent’s life for every year, which would give us a tremendous amount of information regarding social history.

Melinda Mills points out a problem in linking data. There is a huge legal aspect with regard to which data can be linked and which cannot, as participants may have agreed to a way data was going to be used and distributed. Mark Cullen notes that for health volunteers in the health precision initiative, there is breadth of consent, so this would not be a problem and would allow us to study the life course better.

Shripad Tuljapurkar asks how people feel about the idea of forming partnerships with commercial entities? Mark Cullen responds that we should not be so proud to view commercial entities as unworthy partners. Commercial entities attract a lot of talent and have troves of information that are for their own profit-making reasons, but may be useful to us too. Putting aside some of the individual ethical issues, commercial platforms offer extremely valuable data.

Amal Harrati asks to what extent selection plays a part in the new data paradigm. Specifically, health volunteers may have a predilection to health-seeking behavior. Mark Cullen agrees with this sentiment, noting that selection is huge. Dalton Conley notes that although there is criticism about selected samples that are not representative, with post-stratification adjustment, convenience samples can be used more reliably. As long as we know some basic demographics, we can adjust the data accordingly. Ken Wachter elaborates on big data control for selection biases. Post-stratification adjustments have been successful in the field of political science with very strange convenience samples, though sample size here was huge. It’s going to be difficult to perform post-stratification adjustment, unless data linkage becomes more open.

Ben Seligman raises two issues with big data. First, there is the issue of consent. It used to be the case that consent was opt-out. The assumption was that if blood was drawn at an
academic hospital, data would be used for research. Now, consent is opt-in. Should we revisit the issue of opt-out, particularly when data is extended for other purposes than the original purpose? Second, there is the issue of representation. We don’t care so much that data has to be equally representative, we care that there is just enough data to be representative of an individual population. In other words, we don’t necessarily want representativeness, we want representation.

Amal Harrati emphasizes that selection and representativeness are two different issues. The samples matter a lot. When we think about effect sizes, who’s coming to the sample to begin with, and the differentiation within these samples that lead to a proposed effect, we need to be cognizant of how a sample is selected. Daniel Belsky puts this in more concrete terms. In longevity research, this is an extreme problem. What determines variation in lifespan in individuals older and younger individuals may be different. What we uncover may be tailored to populations that we don’t fully understand. Those in treatment who also volunteer their data may be a unique population in itself.

Shripad Tuljapurkar brings up two other points with regard to selection and representativeness. In the case of mortality selection, looking at variances between sub-groups could be informative. Looking at variance in age of death between high school dropouts and people with college education reveals that age of death is about the same in the two populations. Adding this component to Ben Domingue’s mortality selection analysis would be interesting to see. Second, in terms of data there can sometimes be representativeness within a population, but not representation of populations outside of it. Ben Domingue states that in that case we can not necessarily translate how genetic markers affect educational attainment in New Zealand populations to U.S. populations.

Ronald Lee asks if it is known how to use a person’s Safeway data. Mark Cullen notes that consumption data from Safeway employees could be really interesting to examine as employees get a 20% discount, so it seems plausible that they buy all their food from Safeway. This would allow us to study medical v. wellness interventions. This study could not be done generally in the population because people have too many choices on where to buy food. Ronald Lee adds that one would want to start by constructing a dozen indicators based on this massive data.

Ken Wachter concludes the discussion by recalling the selection issue. The value of randomization in classical experimental design is not controlling for things we know about, it’s about controlling things about things we don’t know about. While I do agree that it is about representation and not representativeness is key, it does not combat the unobserved variables that are not accounted for in models.

6. Mortality selection in a genetic sample and implications for association studies

Ben Domingue

Introduction

Let’s assume the following hypothetical model: there’s a risk factor and an outcome (e.g. occurrence of a disease). There is a positive association between the risk factor and outcome,
and assuming data is complete, the larger the risk factor, the greater that adverse outcome. However, what is observed is an attenuated outcome. The idea I have about mortality selection is that those high on the outcome may have died, and as a consequence, we have reduced variability an outcome, resulting in underestimation of the effect of risk factor on outcome.

Role of Mortality Selection

The basic problem in mortality selection is that estimates depend on the window of data collection and the time of data collection. Only those whose year of death exceed the year of genotyping are observed. However, the true genetic effect is based on birth year. If mortality selection is independent of genetics and outcome, perhaps development takes place after year of genotyping minus birth year (e.g. examining Alzheimer’s in a young population). If development is complete by age of genotyping - birth year (e.g. smoking in older cohorts), then we may observe a bias, in particular, a reduced variation in outcome. Here, the observed association would be less than the true association.

Data

Data from the Health and Retirement Study (HRS) was used for this study. The HRS started in 1992, and genotyping in the project began in 2006 and continued in 2008. There are two steps needed to be in the genetic sample in HRS: 1) Mortality selection - participants needed to be alive in 2006 and 2008, but 7,000 of the original 37,319 participants had died by that point and 2) Participants had to be around, available, and elect to participate in the genetic sample. Ultimately 12,000 participants were left. While both steps have implications for how long one lives, we focus on mortality selection. The idea is that mortality selection selects for those are healthier, wealthier, and wiser, meaning that those genotyped should be living longer than those non-genotyped. This is what we see in our data, including when stratifying for sex and race.

So there were two main questions that we tried to address. First, how can we effectively model mortality selection? Second, how might information from these models change inference about genetic associations? Some of the motivating models for examining these questions were the differences we observed in the ratio of means for those who lived past 2006 versus those who died before 2006 (e.g. in education, BMI, and heart disease). Notably, those who lived past 2006 had Alzheimer’s at only about 40% the rate of those who died before 2006.

Results

When modeling early death, you can see that health only predicts death quite well in white males and females. That relationship is strengthened once after accounting for birth year.
To give some context, there has been some literature suggesting that for a heritable disease with perfect information, one would see an AUC of 0.93. Incorporating birth year interaction and random forest variables did not improve the model measurably. We are able to predict as well as we can based on a relatively simple set of factors.

<table>
<thead>
<tr>
<th>AUC</th>
<th>Female White</th>
<th>Male White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health only</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>Health + Birth Year</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>W/ Birth year interaction</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>N</td>
<td>11488</td>
<td>9170</td>
</tr>
</tbody>
</table>

Melinda Mills asks if HRS has data from death certificates related to cause of death, as this could be a good method to examine mortality more specifically. Ben Domingue responds that the data is likely there, but that is something that has not yet been explored.

We can do a pretty good job at minimizing survival differences between genotyped and non-genotyped individuals based on our models. Below are Cox survival curves based on genotype status, age at first interview, birth year and interaction of genotype status for white males and females. The black lines represent those born in 1930, while the red lines represent those born in 1945. The dashed lines represent non-genotyped respondents, and the solid lines represent genotyped respondents. The question is, can we reduce differences in survival between genotyped and non-genotyped respondents using our model for mortality? Visually this would result in same-color curves being closer together in the Cox survival curves.

In particular, I selected those with a high probability of having lived until 2006 based on my model. I adjusted the naïve association estimates for mortality selection such that they become average treatment effects, and did this for a number of outcomes: BMI, height, education, smoking. For all of these, I used polygenic scores. I then looked at static marginal genetic effects (time-invariant) and dynamic effects by birth cohort (time-dependent), and then examined these effects before and after inverse propensity weighting.

Before I discuss these results, I wanted to discuss changes in mean polygenic scores as a function of birth cohort for four different types of outcomes: BMI, education, height, and smoking. You can see that those born in 1920 have a much higher polygenic score for education than those born in 1950. The other three measures show similar results. Clearly, there has not been a long enough time for these changes to be due to evolution or selection, so these trends are presumably due to mortality selection. Looking at the raw marginal associations – the main effects of the polygenic scores on the outcomes, we see that polygenic scores are significantly correlated with BMI, height, education, and smoking.

I wanted to see how these associations changed in magnitude with mortality selection using different types of estimates. With smoking, the marginal effect of smoking is 4-5% smaller after adjusting for mortality selection. More interesting is what happens when considering samples with enhanced mortality selection. We looked at those who had died in the most recent
wave of HRS, and we treated those individuals as if they did not have genetic data. We re-estimated associations, and across all variables, we see changes in the association estimates that are counter to the changes we see after weighting. The fact that these things are in opposite direction relative to the naïve estimates suggests that we really are able to control for the effect of mortality selection.

Conclusion
The first question we wanted to answer was how we can effectively model mortality selection? Our models do a pretty good job of reducing the discrepancy in survival differences between genotyped and non-genotyped individuals. Second, we asked how might this information change inference about certain types of associations? We found that dynamics of polygenic score associations may be biased if not corrected for mortality selection.

Questions/Discussion
Sid Kumar asks how do you determine age of onset of smoking? Ben Domingue responds that in the HRS, they ask people the age of onset, and it generally works. Kathleen Mullan Harris agrees, noting that this is how it is done in Add Health. Daniel Belsky comments that people are pretty good at reporting how many cigarettes they’ve smoked, and that retrospective recall of age of onset is also pretty good. Interestingly, the ever-smoker phenotype seems to be more predictive in GWAS studies, but it is unclear what to make of that prediction.

Shripad Tuljapurkar asks if estimated coefficients can be corrected for in an actual study. Ben Domingue responds that this is exactly what the group is trying to address, but the situation is messy.

Dalton Conley asks about how non-compliance was approached. Ben Domingue responds that people certainly refused to give genetic samples in the HRS. This is a really complicated group, and I’m not trying to model this group. We only have genetic samples and good polygenic scores for white respondents, which is why we only showed associations for these groups. Daniel Belsky adds that less educated, non-white participants in the HRS are less likely to consent to this kind of research. He notes that there may be ways to control for this.

Ken Wachter asks about the differences in the Cox curves and whether they are truly picking up mortality selection. Ben Domingue notes that if you look at the Cox curves, we have survival differences between genotyped and non-genotyped respondents. If conditioned on only those who lived until 2006, you see much smaller differences. The curves would fall within each others’ confidence interval, so what we’re picking up is definitely mortality selection.

7. GCTA and all that

Sid Krishna Kumar

Introduction
The question I’m analyzing today is one of the oldest questions in genetics: what is the percentage of variation in a trait can be explained by the additive genetic contribution. This is commonly phrased as the narrow-sense heritability. In the past, we didn’t have genetic data, so heritability was computed indirectly. In the 2000s, we started to collect a lot of genetic data, and we begin to ask can we compute the heritability directly. When genetic data was first used, we a
large part of the variation in phenotype could not be explained. There appeared to be a missing heritability problem (e.g. 20 SNPs account for only 3% of height). In 2010, a group claimed that they had solved this problem, or at least partially with GCTA. They argued that heritability was not missing, but hidden. When using stringent p-value criteria for genetic variants, many variants that have a moderate effect are left out. Using less conservative criteria, 45% of the variance in height was able to be explained.

Data

The data consists of a matrix of N people and P SNPs. Y represents the N x 1 phenotype vector and Z represents the N x P genotype matrix.

GCTA model

GCTA assumes that every SNP makes a random contribution towards the phenotype. The vector u (P x 1 vector) is describing the contribution of each of these SNPs is. The total contribution of all the SNPs is $\sigma^2$. From these values, variance in genotype, and phenotype can be calculated, and heritability of a trait can be estimated.

$$y = Zu + \epsilon.$$ 

$$u \sim \mathcal{N}(0, \sigma^2 I) \quad \text{and} \quad \epsilon \sim \mathcal{N}(0, \alpha^2 I)$$

$$V_g = P\sigma^2$$

$$V_{phen} = P\sigma^2 + \alpha^2 = 1$$

$$h^2 = \frac{V_g}{V_{phen}} = P\sigma^2$$

Two points that I wanted to make about the model are that 1) this model is precisely specified; it contains all the assumptions needed for analysis. We are assuming that the SNPs are in strong LD with the causal SNPs 2) We have information on N particular people and P particular SNPs, but we’re getting information on a global phenomenon: heritability. What the authors in GCTA noticed is that we are only interested in $P\sigma^2$, so they wanted to estimate this value directly. What they did is they reformulated this expression to get $y = g + \epsilon$, so $h^2 = V_g = P\sigma^2$. The role of P has been completely removed.

What is the insight into doing this? If we knew the k causal SNPs, then the model would run fine. The problem is in reality we don’t only get the causal SNPs. In fact, we don’t even know the SNPs that we’re using, and we don’t know the value of k. To account for this fact, we’re going to assume that the effect of just having the causal SNPs is approximately the effect of having any sort of SNPs. This is the central assumption of GCTA. The question is: does this assumption hold?
The trait being examined here is height, which has a heritability of about 0.6. The blue bars are the estimates for heritability. Now we expect the slope of this line to be about 0.6, but the slope of this line is approximately 0. What this means is that the heritability estimate is essentially independent of the number of SNPs used in the analysis. The red bars here are a correction to the genetic relatedness matrix, and the authors claim that this correction improves the model. After the correction, they find that the slope is almost always 0. This was one of the serious red flags for me. The primary motivation comes from the drop in the below curve from another paper. The authors here took the entire set of SNPs and randomly threw out a bunch of SNPs. As shown below, for the same level number of SNPs thrown away, the heritability estimates are drastically different.

After identifying the problem, I asked myself how this problem could be analyzed. The only input that is going into GCTA is this matrix A. The key thing to note here is that however many SNPs I use in my analysis, the GRM is always going to be NxN. Different values of P are going
to produce slightly different GRMs. What GCTA is assuming is that if I have a set of SNPs and I compute the GRM that is all I need as long as the errors are low. I asked myself: do these errors actually make a difference, and in turns out that they do? How are we going to quantify these changes?

The GCTA equations are shown below:

**LOG-LIKELIHOOD (L) OF THE DATA**

\[
\log P(y_1 | \sigma^2, \sigma^2) = -\frac{N}{2} \log(2\pi) - \log \det(C) - \frac{1}{2} y_1^T C^{-1} y_1
\]

WHERE \( C = \sigma^2 I + V_g A \)

**GCTA COMPUTES THE MLE BY SOLVING**

\[
\frac{\partial L}{\partial \sigma^2} = 0
\]

We show that the second term only depends on the eigenvalues of \( A \), and that the third term depends on the eigenvectors and eigenvalues of \( A \). When we do the analysis, the only thing we’re concerned about is when a different group of people and SNPs are used.

We show that:

\[
\log \det(C) \text{ DEPENDS ON } \sum_{i=1}^{i=N} \log \left( \frac{1}{a_i} + \frac{V_g}{\sigma^2} \right)
\]

\[
y_1^T C^{-1} y_1 \text{ DEPENDS ON } \sum_{i=1}^{i=N} \frac{1}{(\sigma^2 + V_g a_i)} (y_1^T u_i)^2
\]

I’m going to be talking about the case with population stratification. In GCTA the allele frequency of all individuals is assumed to be the same. What happens in natural populations is that they are drawn from a mixture of populations, which makes the eigenvalue distribution of the genetic relatedness matrix \( A \) variable and unstable. The distribution is unstable not simply because there are many small eigenvalues, it is unstable because of the distribution of a few large eigenvalues and many small eigenvalues. This results in the small eigenvalues having a lot of error. In PC analysis, the top few eigenvalues are chosen – you are forcefully setting all the small eigenvalues to 0. In this case, you are not setting anything to 0, so all those errors remain in intact. The eigenvalues in GCTA are packed closely together and there are errors in their values. Ken Wachter asks when the GRM was computed, were the variances of the columns standardized? Kumar confirms that these standardized columns. The gist of the analysis is that small changes in \( A \), result in large changes in the eigenvectors \( a_i \) and \( u_i \), which results in large changes in the maximum likelihood estimate (MLE).
Data
We show that there is possibility for why this can happen, and we demonstrated this numerically. We used the Framingham study, and constructed a genotype matrix consisting of 2,698 people and 49,214 SNPs. The phenotype measured was blood pressure. We then computed the estimate and SE of $\sigma^2(T)$, and created 2,500 sample genotype matrices. From there, we computed the estimate of $\sigma^2(S_1), \sigma^2(S_2), \ldots \sigma^2(S_{2500})$.

Results
The question is what relationship do I expect between $\sigma^2(T)$ and $\sigma^2(S_n)$? We expect the results to be exactly the same because as far as GCTA concerned, we are providing a 2689 x 2689 matrix, which is describing the relationship between those people. We computed the 95% confidence interval using the total population to see if these other estimates are not within the confidence interval. What we found is that most of the distribution lies outside of the 95% confidence interval. The important point is that even if the same set of SNPs are maintained, it is not okay to use different people. Domingue and Boardman found there is consistent heritability even when thousands of SNPs are removed. Even though it looks like the heritability is the same, the number of SNPs is still important. The heritability estimate can be a very misleading quantity to look at.

Questions/Discussion
Dalton Conley asks if any simulations were performed where the linkage between SNPs was varied because this may affect the results? Ben Domingue states that another paper explored this issue, and that GCTA would perform as expected under these different circumstances.

8. Random-effects estimates of heritability for simple and complex traits: some statistical theory

David Steinsaltz

Introduction
Today I’ll be talking about the statistical theory underlying GCTA and related models. This research was inspired by the work that Sid and Tulja did. We came to different conclusions.

We are talking today about a high-dimensional random effects model, which has some interesting properties because of the dimensionality of the model. Let $Z$ be an $n \times p$ genotype matrix, giving the genetic type for each of $n$ individuals at each of $p$ sites, and let’s assume the column variance is 1. Then let $Y$ be an $n \times r$ phenotype matrix, giving observations of $r$ quantitative traits for each individual. For this model, we are assuming $r$ is 1. Finally, let $u$ be a $p \times r$ matrix of genetic random effects (for our purposes just $p$ since $r = 1$), giving the contribution of a unit change in site $j$ to trait $k$. We assumed these to be normally distributed with expectation 0. $\varepsilon$ is an independent noise term giving all the effects that are non-genetic. We write the singular value decomposition of $Z$ as $Z=U\text{diag}(s_j)V^*$. I’ve found it convenient to express this in terms of parameters of $\theta$ and $\varphi$. $\theta$ is the noise precision and $\varphi$ is the ratio of the genetic variance to the noise variance. The normalization...

\textbf{GENES AND COMPLEX HUMAN PHENOTYPES | 23}
constant $p$ gives the contribution of each SNP. $\phi/\theta$ is the total genetic variance. The narrow-sense heritability $h^2$ is equal to $\phi/(1+\phi)$. The way this is analyzed is in terms of a singular value decomposition of the genetic relatedness matrix. In other words, we’re looking for a way of writing it which makes the contributions independent of each other. If I create new variables $z=U^*y$, that gives me $n$ new variables. Now the $z$’s are independent observations with different sizes. They are normal with variances $1+\phi_s^2$. $\phi$, the relative size of the genetic variance, is essentially telling me how different these $z$’s are. The bigger the heritability, the more different they are. If $\phi$ is 0 then there is zero covariance.

We can do maximum likelihood estimation, but this is where I disagree with Sid. I don’t think there is any problem with the maximum likelihood estimation, because the small singular values make small contributions. You can eliminate $\theta$ from the problem, and end up with a profile likelihood: a one variable version of the likelihood. So if we define these two combinations of $\phi$ and the observations

$$w_i = \frac{1}{1+\phi_s^2}, \quad v_i = \frac{z_i^2}{1+\phi_s^2},$$

then this is just saying that $w$ and $v$ have 0 covariance at the solution. If I had the true $\phi$, all these variances would be 1, and that’s why there is 0 covariance. Essentially what we’re doing is we’re trying to find the $\phi$ that matches the variance of these transformed variables $z_i$ with the variance they should have.

So one can compute bias and variance. If we look at the $\phi$ variable, you can get a very large bias if this $\tau^2$, the variance of this $w$ is small, which it often is. On the other hand, if you look at $h^2$, that bias is, in most circumstances, very small.

The GCTA model can be thought of as an oddly formulated linear regression model. If the singular values are close to 1, which in many cases, they will be, then we can approximate it as

$$\log z_i^2 \approx A + h^2 s_i^2 + \log v_i.$$

The model is pretty much unbiased, unless it is truncated at 0. This is because spreads can cover negative values; however, if one assumes heritability has to be positive, the model is truncated at 0 or the negative values are thrown out. If truncated, the negative values are going to have an upward bias. In the original description of the model, $h^2$ was the ratio of the two variances, so the value had to be positive. How you define heritability affects whether it can be positive or negative. In this model, $h^2$ is a really a regression coefficient, representing a correlation rather than a variance. Certainly, there is no reason why it can’t be negative.

Negative heritability would mean that those with similar genotypes have different phenotypes. Is that possible? Based on the following theoretical model, it is. The theoretical model posits that negative heritability can arise from gene-environment interactions. Imagine there are two genotypes: 0 and 1. If you have the 0 type, that doesn’t affect your character. If you have the 1 type, in some circumstances you have a positive character and in others you have a negative character. We can compute the expected square difference – how much you expect them to vary in their character. If effective alleles are uncommon, this will increase in $k$. The overall effect is to make them more dissimilar if they have similar genotypes. Negative heritability has sometimes been reported, but most treat these negative values as statistical artifacts. 

*Spectrum*
The problem is that these are not independent random entries, so the effective number of SNPs is actually much smaller because of linkage disequilibrium. You can get a better fit if you get a smaller value of $P$, which corresponds to the amount of independent information in the SNPs. Dependence between individuals creates small singular values. This explains the saturation effect that occurs in this method. The singular vectors in great generality have a particular distribution, which is like saying they are independent normal except that they have to be orthogonal to each other.

**Conclusion**

- Simple random effects model gives reasonably unbiased estimates of heritability;
- There can be considerable bias if we truncate heritability, but we don't necessarily want to;
- Truncation bias and variance can both get large if there are a large number of independent SNPs.

**Discussion**

Shripad Tuljapurkar comments that there is literature in evolutionary biology where they attempt to estimate the genotypic variance, but when they do the estimates, they often get matrices, which generally do not obey that structure. There are a number of estimations, but lead to matrices that we should not be getting.

Sid Kumar asks: what does the model give you – is it giving the heritability conditional? Does it have a physical meaning? Is it a global parameter? Steinsaltz replies that there is something weird about the normalization. The normalization does depend on which people you happen to get. If you normalize differently, you would get a different variance. You would certainly expect that this to all average out. Sid Kumar then asks: what is the biological interpretation of that number? Steinsaltz responds that the biological interpretation is as follows: having one or the other SNP contributes a certain amount to the phenotype, and that effect is summarized by $u$ when the genotypes are normalized.

**9. Statistical learning approaches to identifying gene-environment interaction**

*David Rehkopf*

**Introduction**

70 years ago, J. B. S. Haldane stated that “the interaction of nature and nurture is one of the central problems of genetics”. Haldane’s work marked one of the earliest attempts to examine gene-environment interactions. Then 25 years ago, Ruth Ottoman at Columbia characterized a number of ways gene and environment may interact, including the classical way we think about gene-environment interactions: the genotype increases expression of the risk factor. In my examples today, we’ll be more thinking about how the genotype modifies the effect of the risk factor, and vice versa.
In the last 10 years, there has been a lot more emphasis on statistical learning approaches to examine epistasis and to a lesser extent GxE. This sort of approach is very interesting. For GxE, we have good priors (polygenic score, our knowledge of the environment and risk factors, etc), but we don’t have great ways to examine GxE. In evaluating the best approach for examining GxE, there are a number of goals that must be kept in mind. The approach must be:

1. Transparent and repeatable: a hallmark with GxE studies is that there has been selective publication of findings.
2. Tractable in small (n~5000) sample sizes: this is important for high quality environmental measures, as meta-analyses may obscure environmental measures and make it difficult to study GxE interactions
3. Robust: we want something that is not spurious
4. Multi-dimensional interaction: a number of genetic and environmental factors may interact together to result in a phenotype
5. Interpretable interactions: this is a problem with a lot of the robust, machine-learning methods because they group genes and environmental factors in ways that are difficult for interpretation

Methods

The approach that I am going to take here is model-based recursive partitioning (decision tree analysis). One adaptation made to this model allows you to fit a structured part of the model, and to search over a number of covariates to see if there is heterogeneity in the relationship between outcome and exposure. In other words, it is not just about predicting outcomes based on a number of factors, but looking whether associations vary, and if there is heterogeneity over sub-groups from any of these factors. We can apply this method to GxE to determine if there are some sub-groups in which the relationship between outcome and exposure is not stable.

A party package in R was used using the “mob” function, which fits a structural part of the model with known confounders and exposure of interest, and then scans over remaining covariates to examine if associations differ by subgroups at each of those cut points. To my knowledge, this sort of method has not been used for genetics studies yet.

Data

Data from the Health and Retirement Study (HRS) was used. The dataset is composed of longitudinal data from those age 50+ and their spouses. Heart disease was the primary outcome measure, and the polygenic score from the CARDIO meta-analysis was retrieved. Included in the model are classical risk factors (e.g. systolic blood pressure, smoking behavior, and cholesterol) as well as environmental risk factors (e.g. education, household earnings, occupation, gender, and age).

Missing data was accounted for using the R package missForest. In our sample, we included only non-Hispanic whites.

We randomly split the sample in half randomly (N=9700), and then created conditional regression trees, as well as nodes and sub-groups for these interactions for half of the sample (training sample). Then, the test sample tested whether these divisions actually held up. For most variables in the model, there are not strong correlations, which is good because if the data are too highly correlated, arbitrary split points can happen during decision tree analysis.
Results
The polygenic score was picked as the primary genetic factor, and highest parental education was picked as the primary social factor. The relationship between polygenic score and heart disease is actually quite linear, so everything else here assumed a linear model. The polygenic score explained about 5% of the variance in heart disease. Parental education had a very similar effect as polygenic score: the relationship was linear and explained about 5% of the variance in heart disease. When adding in the other variables, the two primary variables do not change much.

First regression tree
We then examined whether the relationship between polygenic score and heart disease varied between any two groupings that were created from the decision-tree. The minimum node size is 5% of population to allow for more stable and practical results. The first split point that came out was age (age > 67 vs. age ≤67). The full decision-tree is shown below:

For each of those bottom graphs, on the vertical axis is heart disease; on the horizontal axis is genetic risk score. What these graphs are showing is the variability in the relationship between polygenic score and heart disease by each sub-group. For example, at node 17 we see a strong positive correlation between polygenic score and heart disease in females with a certain genotype between age 72 and 79.9.

I then took the two strongest positive and negative associations, and applied this to the test data to see if the associations by sub-group held up without decision tree-analysis. What we see is that the sub-groups are not dramatically different from the overall estimate, but they are significantly different from one another, as shown below.
Ben Domingue notices that between node 17 and 18, the difference is only in age, but the estimates for heart disease are markedly different and in opposite directions. David Rehkopf comments that these results may be idiosyncratic. On the other hand, there may be something in these results, especially after seeing the results of the test sample. Ronald Lee clarifies that what we see is that in some age groups heart disease is positively correlated with polygenic score, and in other age groups, it is negatively correlated. Ben Domingue asks does this imply that genetics do not matter in females > 80 years old with an earnings <=4.539 thousand per year?

Using a traditional regression tree model that just predicts the outcome, none of the environmental factors nor does the polygenic score come in. This is the motivation for our approach.

Then, using parental education as the main variable in predicting heart disease, I wanted to identify any top hit SNPs that modify that association? The p-values were barely at 0.05, but did pick up some SNPs that had some modification of the relationship. There was not a major difference in the relationship between different nodes. These observations make sense because each SNP does not have that much of an influence on heart disease.

Conclusion

In 3 out of 4 cases, a better social and economic environment was associated with lower polygenic score association with heart disease. The genetic risk was strongest between those aged between 68 and 80. In terms of gender, there was no clear pattern, though 3 out of 5 cases showed weaker associations for men. Finally, smoking behavior was split: 1 out of 2 cases were associated with stronger genetic risk.

I want to come back to the question of the robustness, and how replicable this may be. There have been some great recommendations and proposed guidelines for assessing GxE. There is talk about having very strong priors, which I acknowledge is needed for this work. The approach I have taken is the middle pathway between the complete search and the theoretical
based model for discovery. This is the balance we are trying to achieve. In addition, training and test samples generally support robustness of the approach in this instance.

Questions/Discussion
Ronald Lee asks when one does the decision-tree analysis, is it essentially fitting an additive model, so the risk is higher for people in their 60s than people in their 70s? Could it just be that the true model is multiplicative rather than additive? David Rehkopf notes that this is actually capturing the multiplicative interactions. If there is anything that is causing that relationship to change, it will take that into the account. Ronald Lee suggests that if we wanted to look at factors other than age, one could put age in the structural part of the model. David Rehkopf agrees and says this is one of the things he really likes about the model: one can take certain variables out of the equation and see the effect of other variables.

David Steinsaltz notes that this sort of analysis searches through a huge space of models. He asks how well can an almost exponential models be controlled for? If given enough subgroups, one can find almost any relationship. David Rehkopf says that the solution to this is to deal with sub-groups pretty conservatively. A node can be created only if there are a lot of people in there (>5% of the sample).


Ken Wachter

We’ve had two very different parts of this meeting so far; in the morning, we’ve heard in some depth work that is garnering a lot of attention that is drawing on very large enterprises. This afternoon, we’re hearing a lot that is new. In many cases, this is the first time something is being presented. These issues that may look small and technical have far reach, and what’s being argued is going to affect the direction that this field goes.

In particular, in genome-wide studies, there are 1) GWAS 2) GCTA 3) Polygenic scores. However, one can’t possibly read every paper that uses these methods. The applications have far outrun our understanding of the technical software behind the paper, and statistical side of methods are not really understood. There is a huge impulse to go chasing the next shiny thing.

The fear is that before we understand these methods, we’ll be doing something entirely different. These papers may make substantive claims that may or may not be true because no one has fully understood the methods. This has already happened in the field with twin studies. More and more we hear that maybe the twin studies got it wrong, that they overestimated the heritability due to design problems. We don’t want to do this again. GCTA does lower these estimates, but it is contested if GCTA estimates are still biased upward.

The teams here all agree on the importance of variances when sample sizes are small. I think there is now large agreement that variability in GCTA estimates can be serious in 500,000 size samples. Negative heritability is another important area that I see agreement in. My sense is that most people think to keep the negatives in heritability.

But the big issue in these discussions is how much GCTA is ultimately exploiting population stratification. More specifically, how much of the variability that genome-wide methods are using are reflecting deep ancestry, structurally built-in over thousands of generations, rather than independent samples from homogeneous populations? This issue comes up with polygenic...
scores too - that these scores are not picking up stratification. The singular values in the genotype metrics suggest that there is a lot of deep stratified structure. These issues relate to the problem with twin studies: genetics is confounded with environment. There is a possible confounding between what we are calling genetics and what the genetic markers are telling us about someone’s ancestral history (e.g. it’s not that skin color genes affect anything, it’s the fact that they mark something). The issue of stratification is crucial to consider because it allows us to distinguish whether genetics affect a cellular mechanism, or if they simply mark something exterior.

Dalton Conley comments on the negative perception of twin studies, specifically the reputation of twin studies as self-reproducing and resistant to criticism. Despite all the criticism, there are reports that twin estimates are not that bad. In fact, there are reasons GCTA would give an underestimate of heritability compared to twin models. How did you get a different reading that heritability estimates are overestimated? I came in as a social scientist who wanted to prove heritability studies were overestimates, but failed. Second, GCTA has exploded over the last 5-7 years, but there are other methods that could take off (e.g. assumption-free heritability testing), but haven’t. Why is that the case? Finally, there are two separate issues regarding deep ancestry. I agree we are confounding with ancestry, even after controlling for PCs. With the skin color example, even within families, skin tone varies genetically. This is a form of illegitimate bias, but is a genetic effect, even though it doesn’t strike us as something we want to estimate. That’s different than confounding by ancestry. The reason skin color is an effect is because of the environment of history.

Wachter asks Conley if he can say something about how studies have show a “genetic” predisposition towards rural and urban environments. Conley remarks that you can control for 25 PCs and still see a residual effect and heritabilities that may in fact have a very small genetic component. In the case of urban v. rural living preference for urban living, even if such a predisposition is slightly heritable, we’re finding much bigger GCTA estimates than we would expect. That suggests that there is some bias from deep ancestry.

Daniel Belsky contends that if we’ve learned anything from twin studies, it is that everything is heritable. Predisposition to rural vs. urban living environments could be driven by preferences, traits, and behaviors that are influenced by common genetic variants. Essentially then, there are two stories: 1) genes drive preferences that form environments or 2) accidents of history lead genetically related people to find themselves in environments that are themselves causal of the diseases that we then misattribute to genetics. Ben Domingue remarks that the problem is that both stories can be going on at the same time, making it extremely difficult to distinguish one from the other. Daniel Belsky notes that heritability is not well-estimated. It is contingent on structured variance in the population that is being analyzed, that everything is heritable to a certain extent. If heritability is high, people are suspicious; if it is low, the dependent variable was poorly measured.

Ken Wachter then asks the group what its views of overestimates of heritability are? Kathleen Mullan Harris explains that in Add Health, we started with an embedded genetic design with siblings because that’s all there was. I felt that the heritability estimates were so wide, that you could’ve picked almost anything. What mattered was the types of samples that these estimates were coming from. Up until the time Add Health drew its sample, most samples were not representative and were convenience samples. Over time, as molecular data became available, we were able to test some of the assumptions and realize that the original behavioral genetic models gave us the wrong results (i.e. likely overestimated heritability).
Shripad Tuljapurkar elaborates on this point. Early research had a very cavalier treatment of environmental effects. However, culture frames our perception of environment. For example, it seems reasonable today if you were to have a Wexler test, but there was a lot of debate about this instrument 15 years ago. 25 years from now, we could have a different instrument altogether. Furthermore, there’s a lot of evidence that even with linear, additive models, heritability changes with age – and it can actually even increase with age. The argument is that there should be some strong prior about heritability, and I don’t actually believe this.

David Steinsaltz emphasizes that an important part of the story, which is often forgotten for low heritability traits is that heritability is a ratio. Major non-genetic causes of variability in a trait (e.g. where one lives), could drive that ratio. For example, if there is a cultural influence to live in an urban area where there are jobs, it would be surprising if genetic variants that predispose one to choose a rural environment are going to create a high heritability in this population.

On the other hand, Dalton Conley notes that in a situation in which there is not that much environmental variability, genetic variability would be expressed and predominate. Tuljapurkar remarks that it doesn’t necessarily have to be genetic variability – we can find variability through anything. Going back to the rural vs. urban example, Ben Seligman worries about how variability is going down in the U.S. In terms of where we live, it seems like that is going to greatly confound our estimates. People are choosing to have less children, and secular trends like these must be accounted for.

Kathleen Mullan Harris asks if there is evidence from animal models where one does know the complete ancestry, and as a result, one can manipulate the environment? Shripad Tuljapurkar responds by saying if you look at breeding programs, they do estimate heritability, have decent pedigrees, and have achieved striking results. However, their predictability is poor because they don’t understand why they get correlated responses.

11. Do PCs overcontrol causal effects

*Dalton Conley*

The motivation behind this research stems from the following questions:

1. Are PCs adequate controls for environmental confounding of genetic inheritance in both GREML and PGS analysis?
2. Do PCs over-control for genetic effects that are actually true? If there is a genetic signature that is ancestral, why isn’t that a causal effect? Maybe that is a part of the missing heritability puzzle.
3. Do sibling fixed effects induce bias due to niche formation within families?

*Motivation 1: Is GREML biased by environmental confounding via ancestry*

The key assumption with GREML from the Purcell paper is:

>“Among individuals who are unrelated (i.e. distantly related since all humans are related to some extent), environmental factors are uncorrelated with differences in the degree of genetic relatedness, which themselves result from the randomness of recombination and segregation of alleles below a certain threshold.”
They calculated the relatedness between the matrix of individuals for each chromosome. If their assumption is true, then there should be no correlation between how related two people are on a chromosome because they are sorting independently. They say that none of their p-values are below the Bonferroni threshold, but this is the wrong test. We are not asking if Chr 6 is really correlated with Chr 8 above and beyond a normal distribution, we are asking if there is a weak signal of relatedness between chromosomes. The right test is the Kolmogorov-Smirnov Test where one can examine the entire distribution of p-values. Here then, the fundamental identifying assumption fails.

**Motivation 2: Do PCs over-control for environmental effects?**

**Motivation 3: Is there niche formation within families?**

In other words, if I have a one standard deviation difference between me and my sister in my IQ PGS, does that difference result in a bigger phenotypic difference within the family even with the same common environment than it does with a random, unrelated individual with the same standard deviation difference. The intuition behind this comes from observed effects between siblings being bigger than it is across families. This may have to do with niche formation: families invest more in intelligent child, accentuating small differences over time. I’m particularly interested in this because I am interested in differential investment. PGS gives us a tool to measure differential or compensatory investment in children.

Today we are going to look at four different phenotypes that have polygenic scores, and we are going to go from less environmentally influenced by family to more environmentally influenced by family.

1. Height
2. BMI
3. Education
4. PPVT
5. Depression

For depression, for example, one can image that the second child takes on the sullen role type in response to the cheery first-born. Kids differentiate themselves personality wise and their attitude towards school. This is why depression would result in more niche formation and height not so much. Ben Seligman asks if there is a way of quantifying niche formation? Conley remarks that this is just a hypothesis, but suggestions are welcome.

Data
Data come from the National Longitudinal Study of Adolescent to Adult Health (ADD Health). Specifically, data were retrieved on the 900 sibling pairs.

Results
Below is a graph showing the effect of PGSs across models by phenotype. The vertical axis shows that standardized effect size. The first bar is the effect of the PGS on phenotype with no controls for each of the five phenotypes. The subsequent bars show the effect with PCs, fixed effects, and PCs and fixed effects.
For height and BMI, the effect of polygenic score increases when controlling for PCs, but is pretty stable across all models. Moving to other measures like PPVT (vocabulary test) and depression, controlling for PCs does not change the effect of polygenic score, but the effect goes down dramatically within families.

Conclusion

PCs may be inadequate controls for behavioral outcomes with large within-family variation; on the other hand, height PCs may actually be causal and affect PGS estimation. Change of education PGS across models is dependent on sample, though more careful comparison is needed. Large sample sizes are needed for within-family PC analysis. There should be some sort of consortium formed that pools together data from the Swedish twin registry, the Queensland twins, the Framingham data, Add Health, and the Minnesota data. It would be really interesting to explore these samples for within-family fixed effects, and to have training samples where we estimate the betas from PGS samples from scratch. The combination of PGS and GCTA can be useful for interrogating environmental differences. We can measure variance in genotype and phenotype, but not environment. Using these tools together can tell us something about environmental variance.

Questions/Discussion

Daniel Belsky remarks: you showed some results with 2 PCs and you showed some results with 25 PCs. The heritability estimates were halved when you went to 2 to 25 PCs. How many PCs do you know to control for? There is no real consensus about this number, and no consensus on how to compare PCs. Dalton Conley notes that in this particular case we did 50 PCs because at that point the effect of PCs flattened out. Belsky notes that this point, the point at which an asymptote is reached can be informative. It would give an estimate for how many PCs to control for.

Ken Wachter notes that GCTA is driven by PCs and the eigenvalues of the PCs. It’s not just that we have a category of controls and a category of signals. You can start controlling (i.e. dropping) singular values, and that is going to erode the precision of the estimate, since the signal is giving the variability. Thus, there’s a conceptual problem with removing PCs. David
Steinsaltz adds that by definition the way we normalize these matrices, the singular values squared must average out to 1. It’s not true that we should only care about the big values. What determines the leverage of a component is how far away it is from 1. Somehow you throw away a few PCs and connected to those were 50 small ones, which cumulatively could have a huge effect.

Daniel Belsky remarks that the intuition of taking the first PCs big ones is to use them to control for ancestry variation. The first PCs represent bottlenecks where populations diverged from each other. It is believed that the traits that we study do not similarly arise from coordinated shifts in genome wide allele frequencies. We should see selective pressure around bottlenecks on traits, but that selection should not affect genome-wide patterns of allele frequencies that we are picking up in the first couple PCs, but much subtler shifts in frequencies of allele frequencies.

12. Genetic load and polygenic scores

*Ken Wachter*

Introduction

I want to talk today about genetic load and polygenic scores. I want to start with a quick overview of three parts of what we’re studying: mutation accumulation, functional alleles, and fitness impact. Then, I’m going to show a few results in which we associate ratings from genome-wide evolutionary rate profiling (GERP) with polygenic scores for educational attainment and talk about the concepts involved here and whether we can hope for brighter signals.

Background

We’re going to put three things together, each of which have a model:

1. Genetic load - large numbers of mildly deleterious alleles once held in mutation-selection equilibrium.
   a. We have a model (GCTA) with demographic structure for mutation accumulation
2. List of suspects - SNPs deemed likely to be, or have been, functionally selectively deleterious
   a. We have a model based on genome-wide evolutionary rate profiling (GERP++) ratings
3. Detriments today: complex traits that show effects of sets of SNPs that may serve as proxies for traits with long-term fitness costs
   a. We have model (polygenic scores) to help free us from underpowered single-survey estimates

The idea is something obvious: the SNPs of today derive from the mutations of yesterday. SNPs can fit into one of three stories:

1) Neutral: Mutations which are neutral with respect to fitness. These are mostly pruned out by drift, but can become a permanent fixture in the genome
2) Mildly deleterious: Mutations that are pruned out slowly by selection
3) Potent: Mutations that become either fixated in the genome (favorable) or disappear (unfavorable)
All these SNPs have an evolutionary history.

Model
We have developed a model for mutation accumulation based on mutation selection and recombination, and non-linear age-specific demographic selective costs. Sir Peter Medawar developed a notion of how accumulation of alleles with age-specific effects could shape demographic schedules and account for some of the senescent properties of demographic schedules.
Out of our model come a number of features that can be used for generalizations, in particular, a generalized Haldane principle that allows one to talk about a subset of alleles related to adult mortality to specify a load-free genome and give us bounds on fitness loss. We have something that ties the pieces together, if geneticists can tell us how many functional mildly deleterious alleles people carry. Once we asked that question, we realized that deleterious effects today are based on selective fitness of yesterday, and if that’s the case, maybe we’d see those alleles affecting contemporary health and aging.

The papers seem to say that 50,000 years ago, we should be thinking of effective population sizes of around 10,000. And so, that scale sets the scale for the difference between drift being the driving force behind evolution versus mildly deleterious alleles, which we are interested in. We are looking for alleles that have clearance rates within historical or at least post-agricultural time.

The work of Tennyson et. al (2012) revealed 300-600 functionally deleterious SNPs. These alleles reflect on the properties of aging. There are several ways to categorize functional alleles. GERP++ is one method. At every site GERP++ looks at the alleles of present day, and uses a maximum likelihood estimation, which predicts how many times alleles that were neutral will drift to fixation. This happens enough times so that it can estimate substitutions and changes that happen just because drift. 5.83 is considered a neutral rate. High GERP scores indicate fewer substitutions, and suggest a lot of selective constraint.

Data
Data from the GERP++ analysis totaled over 39 million numbers. Polygenic scores for educational attainment were retrieved from the Consortium of People with European ancestry. A separate GWAS was run on each SNP, so we ran a simple bivariate regression with accompanying beta values and standard errors.

Results
We looked for signal between the polygenic score and GERP rating. If you just take a correlation between GERP rating and polygenic score you don’t see anything. But a large part of the variability in GERP scores is the noise of its maximum likelihood estimation. As a result, GERP scores were restricted to those with a reasonably strong evolutionary rating (GERP > 3 rejected
substitutions). The Betas were split at about the 20% point of the distribution, and placed in a simple two-way table.

More negative Betas are indeed enriched for functionally selective deleterious alleles, but not to a large extent (30 per 1,000 of the functionally computed alleles compared to 24 per 1,000 in the larger Beta group of the functionally computed alleles). Nevertheless, the results are significant: the Chi-Square test rejects a null hypothesis of independence with a p-value of $4 \times 10^{-7}$.

I should note that if split three-way between low, middle, and high Betas, more negative Betas are enriched compared to the middle betas, but more positive betas also seem to be enriched, as shown below. In other words, it looks like what is good and bad for educational attainment are both functionally deleterious. I leave that to your discussion to explicate.

Dalton Conley mentions that it would be great to do GERP++ with Melinda Mills’ data because she’s exploring reproductive fitness and how that relates to genetics. Wachter agrees but notes that he has a different view from Mills, that is, over evolutionary time, fertility was driven by homeostatic response to environment rather than genetics.

Finally, I wanted to touch on one last point. GERP gives a cloudy signal, making it difficult to identify what exactly is a mildly deleterious allele. 5.8 is considered the neutral rating. My trouble, however, is that neutral things ought to have a Poisson distribution. But what is observed is that instead of being enriched with substitutions at the low end, there is a deficit. That worries me about the GERP score. Nonetheless, there are tons of validations in the paper that show GERP scores seem to match well.
Questions/Discussion
Dalton Conley – Nice sigmoid function of the strength of the SNPs and the selective sweep.
Selective sweep is getting much deeper in evolutionary history. There is more time
Tuljapurkar – there is a close connection to selective sweep stuff. The selective sweep literature.
Instead of seeing a positive selective coefficient, you have a negative selective coefficient. What
Ken says is interesting – I would’ve expected more weight on the left.
Wachter – All we want is geneticists to know about these things to tell us how to find functional,
rather than neutral alleles. So looking at selective sweeps if there are enough to do, could be a
way to go. We would be happy to get rid of GERP if we can find better indices.
Lee – selective sweep how many generations do you have mind?
Wachter – People are talking about recent evolution – post-agricultural revolution. 10,000 years;
Lee – something that is a slight advantage – may be too fast?
Belsky – sweep, to my understanding is something big
Zeus (immortal) + salmon (die on reproduction)

13. Genetic heterogeneity in human fertility and educational attainment by country and birth cohort

Melinda Mills

Background
The paper is available to read here: http://biorxiv.org/content/early/2016/04/18/049163.
Genetic factors appear to play a role in fertility. Current genome-wide associate studies
assume that genes for fertility are important for the same countries and regions, but is that really
true? Heritability studies of fertility by country show variation between sexes and countries. This
variation may be attributed to a country’s culture or the specific time period examined in the
data. If we look at age at first birth, there is considerable variation by time and country. It used
to be high in the early 1900s, but dipped by the mid 1940s. There has since been a huge
postponement in age at first birth in developed countries.
Still, why should we examine this? First, GWAS results are prominent in the field, yet the results assume genetic effects on a trait are universal across environments and sexes. Second, examining genetic data by country and birth cohort may help explain the disparity in estimates of reproductive outcomes across twin studies, GCTA, and GWAS. Could this be gene-environment interaction? Could it be that genes explain variance within but not between populations?

To answer these questions, we decided to see if the same genes associated with fertility differ a) across different populations and b) before and after the strong fertility postponement of the 20th century.

Data
Data was retrieved from seven databases across six countries (U.S., U.K., Netherlands, Estonia, Australia, and Sweden). The participants were men aged 50 and above and women aged 45 and above (those with complete reproductive periods), and these participants were divided by cohort (before and after the fertility postponement of the 20th century). Fertility postponement happened at different times in different countries, particularly in Estonia. As a result, it is not the same exact cohort division by country. Number of children ever born (NEB) (N=31,396) and age at first birth (AFB) (N=16,109) were the two phenotypes assessed. We assessed over 800,000 SNPs, and imputed the HapMap3.

Methods
We conducted a mega-analysis using GCTA to estimate SNP-heritability. We fit multiple genetic relatedness matrices by country and cohort, estimated genetic variance, and extended models to look at individuals living in the same population and/or born before or after fertility postponement.

There were four different model specifications:
1. All individuals (genes)
2. Individuals living within the same population (genes X population)
3. Individuals living within the same demographic birth cohort born either before or after fertility postponement (genes X birth cohort)
4. Individuals living in the same population and demographic birth cohort (genes X population X cohort)

The SNP-based heritability estimates was the sum of genetic variance over the total variance in four model specifications.

Results
We see that with gene-population interaction, we get higher SNP-based heritability estimates than from genetics alone. When including genetic, population, and demographic cohorts before and after fertility postponement, we get an almost fivefold increase in heritability estimates for NEB and AFB.
Then, we estimated a bivariate model based on the birth cohort model, which allowed us to investigate whether genetic effects correlated across birth cohorts. SNP-based heritability estimates for NEB within populations were significant for demographic cohorts before and after postponement. Specifically, there was a positive correlation of genetic effects for NEB across demographic period within populations.

Conclusion

We were assuming yesterday no variance in temporal and spatial environments. However, genetic effects on fertility differ across temporal and spatial environments. Geographical and historical environments strongly modify genetic effects associated with human fertility. Taking GxE into account increases the predictive power of heritability estimates almost fivefold compared to the genetics-only model. GxE potentially explains the large discrepancy between individual SNP-based studies and family studies. Finally, we found no evidence for genetic effects shared across environments.

Interestingly, the same pattern is found for education. It suggests that these interactions are important not only in the context of reproductive fitness.

There are a few alternative explanations for our findings. First, there may be heterogeneity in measurement. Across different cohorts for different time periods, definitions differ with regard to fertility measurements (e.g. live births), which affect the way they are measured. Second, there may be population stratification.

Discussion

Daniel Belsky asks: how big were the cell sizes for the stratified data? They say you need at least 5,000 to generate valid estimates. Melinda Mills verifies that the cell sizes were around 4,000, with the exception of Estonia, which was on the lower side.

Dalton Conley notes that when we talk about the fact that PGS R-squared is way below the GREML heritability estimates, we’re comparing apples and oranges because PGS have been coming from consortia that combine across all these different countries and environments, and GREML is collected from around 10,000 individuals from one context/country. I had proposed to Felix to take a single sample like the U.K. Biobank and sub-partition it and conduct PGS and GREML on the different sub-groups. Have you thought about this at all? Melinda Mills notes that the team has started to look at this. Conley remarks that meta-analysis, in general, can be confounding because of the different cohorts and populations involved.

Ronald Lee discusses exogenous factors related to heritability. We could say heritability is only 0.4 because they don’t explain the baby boom and baby bust, but the more interesting
question is how to control for these socioeconomic and cultural factors over time. If we tried to narrow it down to a period and a country, how well are deviations in the population/country at that time explained. I think this is where you’re headed and is exactly right.

David Steinsaltz notes that in this model, the sample was stratified by cohort and country. He asks what is the basic claim for the correlations and numbers shown. Melinda Mills responds that the basic point was to show that different outcomes are seen in populations due to country and time. The next step now is to answer whether genetic effects affect all populations. Steinsaltz suggests that a relevant model may a GWAS across entire sample, but with linear corrections for different categories. This would effectively model the extra variance due to social variables, and show, once these variables are ignored, whether the same genes affect fertility. Shripad Tuljparkar notes that one could start by looking at variance within and between group changes, then look at genetic variation and whether that still affects the outcome.


Nicola Barban

Introduction

Using genetic information can be helpful in assessing the degree of assortative mating on socioeconomic status. A recent study showed that spousal resemblance on educational attainment was very high in the early twentieth century, declined to an all-time low for young couples in the early 1950s, and has increased steadily since then (Mare, 2016). We are interested in the last part: the increase in assortative mating among those with similar educational attainments.

Assortative mating may have direct implications for the transmission of socioeconomic status and inequality across generations (Currie, 2011). In fact, “if matching in 2005 between husbands and wives had been random, instead of the pattern observed in the data, then the Gini coefficient would have fallen from the observed 0.43 to 0.34” (Greenwood et al., 2014). Meanwhile, in many countries women now excel men in terms of participation and success in higher education (Grow and Van Bavel, 2015). We currently do not know how this will affect assortative mating.

When measuring assortative mating it is essential to ensure that the actual degree of assortative is correctly measured, that is, it is neither under- or overestimated. There are a few sources of bias that need to be considered:

1. Measurement error - may underestimate the actual degree of assortative mating
2. Simultaneity bias (reverse causation) - may overestimate assortative mating
3. Omitted variables - may under- or overestimate assortative mating

Omitted variables are what I am most worried about, as matching processes are multidimensional and models need to account for all these dimensions.

Methods

Previous empirical strategies tried to solve these sources of biases, but none of these strategies are perfect. For example, one recent study examined variation in male educational attainment induced by the WWII G.I. Bill (Larsen et al., 2015). However, that study only looked at variation on one side of the marriage market, and was based on cohort variation. We contribute to the literature in two major ways. First, we use an instrumental variable (IV)
approach based on spousal polygenic scores of educational attainment to quantify assortative mating in the marriage market. Second, this method allows us to assess whether assortative mating in the marriage market takes place alongside additional characteristics correlated with education.

**Measuring assortative mating**

Suppose that we have two populations of equal size, normalized to one. Agents differ in their educational attainment. The stochastic matching functions are given by the following equation:

\[ y = \alpha + \beta x + \nu_y \quad \text{and} \quad x = \alpha' + \beta' y + \nu_x \]

where \( x \) denotes the educational attainment for men, \( y \) denotes the educational attainment for women, and \( \nu_y \) and \( \nu_x \) are random components. With a valid instrument \( Z_x \) for \( x \) and \( Z_y \) for \( y \), we can estimate the betas, the degree to which educational attainment of one spouse can affect the other:

\[ \beta_{IV} = \frac{\text{cov}(y, Z_x)}{\text{cov}(x, Z_x)} \quad \text{and} \quad \beta'_{IV} = \frac{\text{cov}(x, Z_y)}{\text{cov}(y, Z_y)} \]

We know from previous literature that this is very difficult to model because there are violations of core assumptions that need to be made. The most important assumption is the exclusion restriction assumption: that spousal polygenic score for education affects one’s own education only through spousal education, net of own polygenic score. To satisfy this assumption, instead of using one’s own polygenic score to predict one’s education, we use the polygenic score of the individual’s spouse to predict educational attainment.

Dalton Conley asks if it has been considered whether spousal education is complete by the time of marriage? If education is not complete (e.g. for many that are pursuing doctorate-level education), then this would violate the exclusion restriction. Nicola Barban acknowledges that education is examined only at the time of marriage; the assumption here is that assortative mating is done based on the time of education. Conley contends that people plan together for their future education and that the sample should be isolated to people who met after they both completed their education, otherwise the exclusion restriction assumption is violated. Nicola Barban agrees with this suggestion. Dalton Conley admits that he is a skeptic for using genes as IVs. There needs to be a placebo test and a pathway that works through the environment in this population, and another population where that environmental condition does not hold to show that there is no reduced effect. Nicola Barban says that the novel thing with this study is using one’s spousal genetic data to predict educational attainment. Furthermore, even if the exclusion restriction assumption does not hold, it can still provide valuable insight in assortative mating.

Sid Kumar asks if a strong signal is expected, given that the approach is indirect (using spousal polygenic score as a predictor). Barban notes that other instruments, such as quarter of birth models, were tested, but this model was the most powerful.

**Data**

Data from the Health Retirement Study (HRS) was used. Only white heterosexual couples at their first marriage were included (N=2,886 individuals or 1,443 couples). We controlled for place of birth, year of birth, an indicator variable if the place of birth differed between spouses, and the first five PCAs.
Results

The correlation between a wife’s and husband’s years of education was 0.56. The correlation between a wife’s and husband’s polygenic score was 0.16. The regression results using wife’s and husband’s polygenic scores as instrumental variables are shown below:

**Using the Wife’s Polygenic Score as an Instrumental Variable for Wife’s Education**

<table>
<thead>
<tr>
<th>Years of Education</th>
<th>Wife</th>
<th>Husband</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FS</td>
<td>OLS</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Wife’s Years of Education</td>
<td>0.630*** (0.030)</td>
<td>0.783*** (0.188)</td>
</tr>
<tr>
<td>Husband’s Years of Education Polygenic Score</td>
<td>0.347*** (0.063)</td>
<td>0.280*** (0.066)</td>
</tr>
<tr>
<td>Wife’s Years of Education Polygenic Score</td>
<td>0.385*** (0.063)</td>
<td>0.059 (0.070)</td>
</tr>
</tbody>
</table>

*F*-test instrument relevance 36.93*** – –
Hausman test – – 0.64
Hausman test p-value – – 0.423 [0.423]
Observations 1,419 1,419 1,419

Bootstrapped standard errors in parentheses. Bootstrap Hausman test is reported.
*** p < 0.01, ** p < 0.05, * p < 0.1

**Using the Husband’s Polygenic Score as an Instrumental Variable for Husband’s Education**

<table>
<thead>
<tr>
<th>Years of Education</th>
<th>Husband</th>
<th>Wife</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FS</td>
<td>OLS</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Husband’s Years of Education</td>
<td>0.431*** (0.022)</td>
<td>0.692*** (0.113)</td>
</tr>
<tr>
<td>Wife’s Years of Education Polygenic Score</td>
<td>0.301*** (0.079)</td>
<td>0.256*** (0.055)</td>
</tr>
<tr>
<td>Husband’s Years of Education Polygenic Score</td>
<td>0.508** (0.078)</td>
<td>0.059 (0.070)</td>
</tr>
</tbody>
</table>

*F*-test instrument relevance 42.70*** – –
Hausman test – – 4.04**
Hausman test p-value – – 0.044 [0.044]
R-squared – 0.34 –
Observations 1,419 1,419 1,419

Note: Control variables are described in Tables 3 and 4.
Bootstrapped standard errors in parentheses. Bootstrap Hausman test is reported.
*** p < 0.01, ** p < 0.05, * p < 0.1

Conclusions
Genes and Complex Human Phenotypes | 43

Assortative mating on education is typically underestimated. The ratio of the estimated OLS-IV coefficient for wife’s education is 0.81 (underestimate of 20%), and for husband’s education is 0.63. We reject the hypothesis that OLS and IV coefficients on husband’s year of education are the same. The fact that they are different tells us that there is an important role of gender in socioeconomic attractiveness in the marriage market. Women may place more value on other male characteristics, such as income.

Discussion

Dalton Conley comments that the IV strategy can be used for two reasons: 1) to purge measurement error or 2) to identify causal effect. I don’t know what problem is being solved here by using IVs. Can the difference between men and women be that there is more measurement error in men or women in years of schooling? Second, regarding the causal effect, if I increase my years of schooling by one year, it’s not necessarily causing my wife’s schooling to increase by x amount. Nicola Barban responds that the study gives the real degree to which assortative mating occurs. Sid Kumar notes that there is no indication of causation in this model anyway. These are all correlations.

Wachter asks what’s your take on genetic homophily between spouses? How does that affect what you’re doing? Nicola Barban responds: there is some similarity in spouses, but I’m a bit more skeptical on genetic homogamy between spouses.

Daniel Belsky asks if the basic observation here is that genetic variance in the husband’s education is the attractive thing, more so than the observed education? Do women value men who are smart and hard-working (partially genetic traits), versus those who have fancy degrees? It just happens that these men may have fancy degrees as well. Nicola Barban says that that may be true, but people don’t currently know too much about the genetics of educational attainment, and education may serve as a good proxy for those traits.

15. Implications of the research for social science

Jeremy Freese

It is obvious that the field of social science and genetics is making clear progress of a certain type. One wonders at the same time if there is not progress of other types. Another way of putting it is that there are enormous advances in data availability and technical apparatuses; however, in some conceptual respects, I worry that the field is a bit like a blindfolded child that has been handed a bat in a room that may or may not contain a piñata. To what extent does the field really know what it is doing?

We use PGS as a tool, but do we truly know what this number represents? In the case of educational attainment, we recognize that the PGS is not a score of education per se. It is a total score of what we deem to be the genetic predictors of education. Other predictors are excluded, and controls for principal components are in place, but we don’t know what and how many variables to control for. Why do we include principal components? Is it confounding? Can it exacerbate or attenuate causal inference? I am interested in these questions and the views others hold on these topics. As a spectator, I am confused about what we apprehend to be the units of the score, and how we determine what units should be included in that score. If we don’t have a clear idea on the score, we can produce all sorts of findings that may or may not be accurate.
Daniel Belsky remarks regarding the principal component issue that, when looking at variables empirically, some are much more highly correlated with the largest principal components than others. Educational attainment scores are not highly correlated with PCs, and the PGS does not vary substantially with more PCs. But between populations that are different, we are concerned that other differences between populations confound results and PCs matter. We don’t have a great solution, but we do have PCs that track for population differences.

Ron Lee asks for an example of what traits may be excluded from PC analysis. Belsky responds with a hypothetical example. If people in Scotland eat less healthily than those in France, independent of genotype and due entirely to accumulated cultural tradition, these principal components would account for those differences. In essence, we’re trying to remove genetic variants that may be correlated but are not causal to the phenotype of interest.

David Steinsaltz asks about phenotypes that are reflected in the genes, but for non-genetic reasons. Daniel Belsky gives the example of the schizophrenia polygenic score, which gives high weighting to the major histocompatibility complex of genes on Chr 6. These genes are involved in immune function, and this region of the genome looks very different in different human populations because these populations have adapted to various microbial environments. It could be the case that those microbial adaptations are causal of schizophrenia in some way, but absent evidence, we would like to exclude that variance from our genetic prediction of schizophrenia. It is a conservative assumption, but we don’t want to assume that population differences in genotype are causal in population differences in phenotype until we see evidence that within a population the genotypic differences predict the phenotypic differences.

Ben Domingue notes that with GWAS studies it could be extremely detrimental to identify a false positive that result in the development of expensive, but ineffective gene therapies. Dalton Conley remarks that when examining the first two principal components of racial identity in the United States, they predict race almost perfectly. However, you don’t want to say that these associations are genetically causal when in fact it may have to do with cultural factors or discrimination. At the same time, PC1 and PC2 map really cleanly actually in the Netherlands. Humans seem to be sorted genetically even at seemingly the most micro of scales. Shripad Tuljapurkar finds PCs very hard to interpret in a geographic context, and notes that some people don’t even use GWAS and PCs to interpret this kind of data.

Ken Wachter remarks that one of the things from statisticians that comes out very strongly is that each of these PCs can be compared against very closed form projections from an independent setting. There is a standard against which to see how much these variances differ from an independent case. There are a number of singular values where there are huge differences. I assumed I could look at a geneticists’ model for what’s the shape of those genome matrices that one should expect, but when I went to find these papers, I couldn’t find that. Wachter remarks that he is baffled by the huge amount of effort that has gone in coalescent studies, and yet, there still isn’t a model for this stratification. There is all this information now about the origins of the genome. Why can’t we look up a model for what the singular values of a genotype matrix with well spaced SNPs should be?

Jeremy Freese remarks that there are different types of inquiry that are making use of the polygenic score. You can imagine that there is a certain optimum, and the further you get away from that optimum, the smaller the influence of genetics. I’m not sure what tool there isto fully assess the optimum. More concretely, one can imagine that there is an age distribution optimum underlying the development of the educational attainment score. Educational attainment could be relatively optimized for some age groups compared to other age groups, and the further away...
from that optimum, the smaller the influence of genetics might be. Daniel Belsky makes the distinction that it is the increasing or decreasing variance of those genetics that are particular to the study population. This is the difference between variance decomposition and polygenic scores. The former is a precise prediction of how a suite of genotypes relates to a phenotype. It’s true that this precise prediction is optimized for the study population. I don’t think that distance is spurious. The fact that correlation gets smaller tells us that genetics are less important in that population.

Freese notes that two possibilities exist. The first is if a GWAS was developed with respect to one group, and that one group is much more representative in developmental score than another group. The second is if there is something more stochastic about educational attainment that may not be related to genotypes, but will also look like a weaker signal in a different population. In either case, we need to assess what we know. Do we have a good sense of what it means to see a higher score in one population versus another population? Belsky notes that as long as we know that population, it is okay. Imagine a regime where education is encouraged versus a regime where education is discouraged. That may influence how educational attainment genes are reflected and the extent to which they are reflected. We would also learn about how social and political policies mediate the relationship between genetics and educational attainment.

Dalton Conley notes that in studies that use many cohorts, there is a disadvantage of not understanding specific populations as well. However, the advantage is that the signals we do get are biologically robust and these signals should not vary by environment or population as much. That is a problem for GxE work using many cohorts because the environment has been estimated across the range of modern environments based on common signals. Belsky notes that there are exceptions depending on the scope of a study. If you look at the Alzheimer’s studies predicting cognitive decline, they are careful to select only brains that have validated Alzheimer’s pathology, so it may not be that representative of a sample.

Ronald Lee recalls Ken Wachter’s discussion on GERP, and revisits the topic of why some genes are under selection and others are not. The question of fitness came up yesterday in the context of AFB and NEB. Fitness is operationalized as the ability to grow rapidly in numbers, more rapidly than rival gene lines. Alternatively, it is the ability to equilibrate at higher density than other gene lines, or better survive environmental variation. Fitness can be seen as flowing through the demographic variable, so that fertility and mortality appear primary to the variable. This can sometimes be misleading because the success of a hunter-gatherer group depends on the direct genetic effect on fertility, yes, but also a massive indirect effect: the ability to get food, digest food, think, and negotiate. In the context of genetic sweeps, there is this idea that there have been major evolutionarily movements in cognition in the last 1,000 years. Now these kinds of traits that do not have so much to do with deleterious mutations, but more with positive selection, they probably involve tradeoffs. For example, larger would use more energy. If we have an optimizing framework, there will certainly be deselection of deleterious mutations, but also the positive selection that gets you to your moderate self. It’s not so obvious what to look for in GERP; you can imagine that there are traits for which there are both positive and negative variations that would be selected against.

Shripad Tuljapurkar notes that genes may hold little predictive power except in very restrictive settings. We are not looking for genes that tell one whether they are going to be taller or shorter. That’s the bias we have because a lot of the modern technology in genetics has been driven by medical science. What does that leave you with? It’s really all about variance. I think
there are multiple optimal life histories that exist. You might have variation in certain genes, so it’s informative. You really are looking at variation. It’s not surprising you see variation around a single optimum because not everyone is optimal. There are many processes that push individuals away from optimal. It is a non-equilibrium process. You are seeing a lot of things in motion. Maybe we need to think more about variance decomposition. When we look at things we look at disparities. How much of that disparity do we need, and how much can we change purely by changing environment, and how much of that disparity due to genetics can we minimize?

David Steinsaltz asks why we use linear models? There is the argument that if there were multiplicative effects you would get an explosion of variability of when things recombine. But of course that is an even stronger argument for the opposite, which is this notion of canalization, that in fact, effects are sub-linear. The cumulative effects of multiple changes tend to in fact be less. Some of the anomalous results we get when we try to combine samples and looking over large number of SNPs may reflect that. What seems to be a zero effect SNP may be positive in combination with some things and negative in combination with other things. We don’t have the right vocabulary for what it is we are hoping to discover by looking at heritability. Heritability is so sensitive to environment. When I look at heritability, what I am interested in is the story behind it. The analyses that are centered around trying to figure out how much of the variance is genetic is far less interesting than the story and mechanisms behind it. The real goal should be to unravel that story. Ken Wachter agrees that mechanisms are important, but does note that small things can hold large difference over time, and can be insightful for exploring evolutionary mechanisms.

16. Development of a measure of cognitive age

Amal Harrati

Introduction

Instead of focusing on a genotype today, I’ll be focusing on a phenotype. And specifically, I want to talk about the development of a measure of cognitive age, particularly at old age. The goal for my research was to create a composite measure of cognitive function that is stable and can be replicated between individuals, over time, and across surveys.

Motivations

Development of a measure for cognitive age is important for a few reasons. We know that dementia is a growing and important health problem. More than 35.6 million people live with dementia worldwide, a number that is projected to increase to 115.4 million by 2050. Age and genetics are two of the biggest risk factors for cognitive decline and dementia. When we talk about dementia, we’re talking about all sorts of things. We’re talking about a big convolution of things that include normal cognitive decline, and Alzheimer’s. The attempt here was to examine middle-aged function as a summary measure, and examine how we think about declines. However, on both the phenotypic and genotypic side, measures of cognitive function are often noisy and unstable, particularly across time. There are a variety of statistical techniques to force things to look a certain way, but we wanted to develop a summary measure that avoids this pitfall.

Methods
Cognitive function can be measured a variety of different ways based on the different aspects of cognition (e.g. crystallized v. fluid intelligence). Cognitive age, like other measures, captures one of many aspects of cognitive function, but our hope is that this measure is more stable.

We are borrowing here from the biological age model, which, in some circles, has a bad reputation. Part of the problem with biological age is the marketing. Biological age is marketed as a true global state, this broad idea that could capture all of aging with one measure. We are marketing cognitive age as a summary measure that is meant to capture a series of cognitive functions, rather than a true global state, which does not necessarily exist. The basic idea is that we have a set of age-dependent variables, usually something like biomarkers, and some kind of algorithm for evaluating these variables.

We used the Doubal and Klemera algorithm that took some of the current algorithms of biological age and extended them a bit, resulting in a few nice properties. The algorithm has four basic assumptions that are important to discuss.

1. Aging is a natural process which is manifested in changes of many properties of living organisms. The speed of the process is markedly different for various species, and to a much more limited extent, also for individuals of the same species including humans
2. Differences in biological age, as far as people of mutually equal chronological age are concerned, correspond to the differences in their individual degree of aging
3. Any measurable property of human organisms that changes systematically with chronological age might be affected by the individual degree of aging and can then be used as biomarkers
4. Long-range course of any true biomarker is governed by the individual’s biological age, but it’s actual value X can also be affected by biological age-independent transient random effects, by fluctuations.

Below is the basic formula for cognitive age, where $x_{ij}$ is the individual-level value of cognition tests, $q_j$ is the intercept of regression of cognitive test j, $k_j$ is the slope of regression of cognitive test j, $s_j^2$ is the root mean squared error, and CA is the chronological age:

$$\sum_{j=1}^{m} (x_{ij} - q_j) \frac{k_j}{s_j^2} + \frac{CA_i}{S_{BA}^2}$$

$$\sum_{j=1}^{m} \left( \frac{k_j}{s_j} \right)^2 + \frac{1}{S_{BA}^2}$$

Essentially what we are doing is that for any individual-level cognition test j, we measure the individual value. For every cognitive test j, we regress on age, and save the intercepts, slopes, and root mean squared error. The calculation of cognitive age is ultimately composed of a battery of cognitive tests, based on a set of individual-level cognition test j.

Data
Data came from the Health and Retirement Study (HRS), which is actually an exceptional dataset for measuring cognition, particularly as people age. As a result, there are a variety of cognitive tests and measures at the individual level that are asked over a long period of time. Some genetic data was also used, specifically on APOE, a driver of Alzheimer’s, and genetic risk scores. Data from Waves III-X (1996-2010) were used (N=25,449).

Four individual measures of cognition in the HRS were isolated:
1. Immediate recall test: 10 words were read; respondents were asked to recall as many as possible immediately after the test
2. Delayed recall test: same as above, but with a 5-minute delay after the test
3. Serial 7 test: respondents were asked to count backwards from 100 in intervals of 7
4. Backwards counting: respondents were asked to count backwards from 20 to 10 for 5 trials

In addition, there are two cognitive-related summary measures in the HRS:
1. Mental status summary score: an aggregation of the four individual scores above and a few other cognitive scores
2. Total cognition summary score: an aggregation of all cognitive-related scores

Tuljapurkar asks how these tests are scored. For the summary scores, the results of the individual tests are simply added up.

Results
For the remainder of my time, I wanted to accomplish three things:
1) Show what the data look like
2) Illustrate that the data is reasonably associated with what we do know about the genetics of dementia
3) Compare cognitive age as a predictor of cognitive function to other measures

Looking at a scatterplot of chronological age v. cognitive age, one can see that, as expected, chronological age and cognitive age are highly correlated, but do show a considerable amount of variation between individuals. Next, looking at density plots of cognitive age gap by age group, one can see differences in age gap by age group. On the x-axis is chronological age minus cognitive age, and on the y-axis is density. Here, a number > 0 is “better”, that is to say, someone could have a chronological age of 65, but a cognitive age of 60, resulting in a value of 5 on the x-axis. The plot is further divided by age group: the red line represents those aged 50-60, while the blue line represents those aged 70-80. Between age groups, the density plots are not widely disparate.
Then, a simple model on sex confirms what we already know about cognitive age: in general, males have a higher cognitive age than females, and non-Hispanic white individuals have a lower cognitive age than non-Hispanic black individuals. Main effects of education also are consistent with the broader literature: those with higher educational attainment exhibit a lower cognitive age.

We then examined the genetics behind cognitive age. We looked at individuals with 0, 1, or 2 alleles for the E4 APOE gene. In the graph below, negative numbers represent a lower cognitive age for a given chronological age, while positive numbers represent a higher cognitive age for a given chronological age. Interestingly, there is no large gap between chronological and cognitive age in individuals with no E4 alleles. However, for those homozygous for E4, individuals seem to have a higher cognitive age – a worse gap - than what is predicted by their chronological age.
Ronald Lee asks if this is a linear model, and if so, if a non-linear model has been considered to examine the data. Harrati confirms that this is a linear model, but that analyses are preliminary, leaving open the idea of exploring this data with non-linear models.

Next, we wanted to look at simple correlation scores using two existing polygenic risk scores on cognitive function. The goal here was to see how our cognitive age measure compared to other similar measures of cognition. Correlations were small across the board, which is part of the challenge with dementia data – it is underpowered. Still, all correlations were in the expected direction. Notably, correlations for cognitive age gap did better or just as well as other measures of cognitive age.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Marden PGS</th>
<th>IGAP Risk Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Age Gap</td>
<td>-0.0567</td>
<td>-0.0450</td>
</tr>
<tr>
<td>Self Report Memory</td>
<td>0.0113</td>
<td>0.02608</td>
</tr>
<tr>
<td>Self Report Mem Past</td>
<td>0.0194</td>
<td>0.0008</td>
</tr>
<tr>
<td>Dementia/AZ</td>
<td>0.0574</td>
<td>-0.0025</td>
</tr>
<tr>
<td>Memory Summary Score</td>
<td>-0.0532</td>
<td>-0.06225</td>
</tr>
</tbody>
</table>

Finally, binary regressions were performed comparing cognitive age, chronological age, and the summary score to actual health outcomes such as self-reported memory and self-reported health. This is really a first stab at the data – there are no controls – to see to what extent do effect sizes look reasonable and what R-squared is telling us. Please note that the health outcomes are reverse-coded, so a low self-reported memory score would correspond with a high cognitive age,
and vice versa. From these correlations, it appears cognitive age is a better predictor than chronological age for these five health outcomes.

<table>
<thead>
<tr>
<th></th>
<th>COG AGE</th>
<th>AGE</th>
<th>Summary Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimate</strong></td>
<td><strong>R²</strong></td>
<td><strong>Estimate</strong></td>
<td><strong>R²</strong></td>
</tr>
<tr>
<td>Self-report memory</td>
<td>0.0194</td>
<td>0.0655</td>
<td>0.0136</td>
</tr>
<tr>
<td>Self-report past mem</td>
<td>0.00454</td>
<td>0.0139</td>
<td>0.00635</td>
</tr>
<tr>
<td>Self-reported health</td>
<td>0.01862</td>
<td>0.0368</td>
<td>0.01358</td>
</tr>
<tr>
<td>How old do you feel?</td>
<td>0.865</td>
<td>0.4043</td>
<td>0.78135</td>
</tr>
<tr>
<td>Prob of work after 65</td>
<td>-0.9532</td>
<td>0.0266</td>
<td>-0.30086</td>
</tr>
</tbody>
</table>

Conclusion

Using borrowed constructs from biological age, cognitive age provides a stable and intuitive measure of cognitive function. Cognitive age is associated with a number of socio-demographic variables. It is also associated with known genetic risk factors of dementia and memory loss. Finally, cognitive age outperforms age or other summary cognitive measures in predicting a number of health and memory outcomes.

Discussion

David Steinsaltz asks if the time of day of doing cognitive tests might affect evaluation of cognitive decline. A few studies suggest that if individuals are measured at their optimal time of day, a lot of the trend in cognitive decline disappears. Older people also have been shown to be more focused in the morning. Harrati states that she has yet to be convinced that the way that we are surveying (e.g. survey time of day) may entirely wash away the observed trends.

Steinsaltz then remarks regarding the formula used to calculate cognitive age that one is essentially taking a weighted average of various measures. Why did you decide to do it this way, and not take into account the correlations among the different variables? Harrati responds that aside from the correlation between immediate and delayed recall, across measures, the correlations are quite low, as they are pretty distinct concepts. However, thinking about the extent to which the two recall measures are highly correlated is something we need to spend more time thinking about. Daniel Belsky adds that, empirically, in predicting mortality, Harrati’s model works better, as correlations between variables are below the 0.5 level.

Belsky discusses how as you become chronologically older, you become more special due to mortality selection. As a result, cognitive function, as presented in this model, may not be completely linear. It would be interesting to apply Ben Domingue's mortality selection model to this kind of data. The advantage of this measure would be that it would give older adults a metric for how they can function in the real world. Ronald Lee notes that this or other minor transformations could be great for comparing cognitive aging in a variety of surveys. Shripad Tuljapurkur also emphasizes the importance of speed in the cognitive aging discussion. What is meant by losing cognitive function faster? Amal Harrati responds that to lose cognitive function
faster is to proceed faster than the normal variation, based on chronological age. It is comparing the norm of the sample to that individual.

17. Discussion: The aging process – how do we measure it in this setting? How does it influence results derived from genetic studies?

David Rehkopf

There is no question we should re-weight for those who select into genetic samples. Weighting should account for those people who leave a sample not due to death, as these individuals could have special circumstances. There is also the interesting etiological question of whether we should weight for mortality selection and how much we should weight for it. In population health, it is generally agreed that it should not matter who has already died because we care about those who are living. However, in considering genetics, should we weight and account for mortality selection?

Ben Domingue contends that, with genetics, to understand an issue such as smoking persistence, it is critical to account for mortality and to re-weight accordingly. Otherwise, there it would be problematic to infer results based only on the living subset of the population. Daniel Belsky agrees and adds that such re-weighting is necessary to make inferences about the general population, as selective non-participation due to mortality can influence the results of study. He mentions a GWAS study that identified genetic variants related to exceptional smokers – those smokers age 90+. People who carry exceptional smoker genetic variants do seem to live longer and are less susceptible to cancer mortality. Ronald Lee notes that to see if those results hold generally across the population, re-weighting must be done.

Daniel Belsky notes an interesting conundrum from David Rehkopf’s presentation from yesterday: a higher genetic load for heart disease is actually protective in old people. However, this result may be attributed to special protective factors that these old individuals possess that prevent them from dying from heart disease. Those without those protective factors would die at a much earlier age, which would underline the need for mortality reweighting.

David Steinsaltz adds that a number of risk factors of various sorts show a similar pattern: high diastolic blood pressure is associated with lower mortality in people aged over 75. It is also well-known that higher weight is associated with lower mortality in elderly populations. One possible explanation for this phenomenon is that the physiological effects that cause people to die vary by age. Daniel Belsky adds that optimal levels of biomarker (e.g. blood lipids) may change with age, which should influence the way studies are designed, and interpreted. Another explanation is that these may be selective effects, in which case mortality reweighting should be done.

Ronald Lee comments that one thing is certain – analysts and researchers must make a conscious effort to design their studies and present their results. They ought to be able to explain their method, and perhaps present both ways: weighting for mortality, and not weighting for mortality. There is no right or wrong, necessarily, but the choice must be justified.

David Rehkopf remarks that weights are rarely used in the field currently; the takeaway from this discussion is that there should be a lot more re-weighting done, even for studies that have already been conducted. Melinda Mills agrees, stating that there has not been much of an
impetus for reweighting despite all the talk in the field about having a representative sample. Ben Seligman notes, however, that reweighting is not always feasible. In the Health and Retirement Study (HRS), re-weighting can be done, as genotypes are collected and individuals are followed across time. With the UK Biobank, however, such reweighting is not as feasible. Rehkopf notes that reweighting could still be accomplished by using Census data.

Kathleen Mulan Harris asks if there has been an effort to characterize all GWAS studies. Such insight would be informative because it would reveal how many GWAS were performed on specific populations, for example, adolescents or adults. Currently, there appears to be a skew towards older ages. Shripad Tuljapurkur raises an additional question: are there factors that we do not know, which are causing these people to die? Such factors need to be examined in order to accurately weight for mortality.

David Rehkopf asks what are the motivations for creating age metrics, and what is the utility of the composite measures of aging. Daniel Belsky remarks that he likes composite polygenic scores because by lumping multiple tests together, the measurements that result are often more desirable. Aging, after all, is a biological phenomenon that affects multiple systems in the body in a coordinated manner. As a result, measurements should reflect the effects of multiple systems.

Chronological age – we are not interested in this variable. We are interested in how much time they have left; predicting function and functional decline in this sense would be more important

Amal Harrati notes that it is strange to emphasize precision on the phenotypic end, but composite measures on the genotypic end. There seems to be a prevailing idea that we can’t aggregated based on phenotype, but can aggregate based on genotype. Ben Seligman notes that this may come down to a matter of measurement. Genotypes can be easily aggregated because they are on the same scale, but phenotypes cannot readily be aggregated. David Steinsaltz disagrees that genotype can be measured on the same scale.

- Seligman: aggregate genotype – they are all on the scale; but how do you aggregate the phenotype
- Steinsaltz: don’t really agree that genotypes are measured on the same scale. We’re talking about differences in genotypes between people – and the differences depend on the frequency of the SNP. Well we normalize, but assume that all SNPs have the same variance
- Steinsaltz: The way these combinations work- you’re not taking bone density and averaging it with memory score; you’ve got the average bone density of an 80 year old and trying to put all those ages in an appropriate way

Shripad Tuljapurkur ends the discussion with the following words. In the end, I don’t think we’ll ever get an answer if we rely on the data because it’s a moving target, and we keep changing, and as soon as we learn something, we run out and we tell people to do something different. Maybe we need to ask a different question: what are the most effective interventions to recommend 40-50 year olds. And what is effective? We’re about to see the greatest transformation in human history – you’re going to want to people aspiring to a much higher expectation for how they function because they expect more. We want to bring an informed,
scientific perspective in that. But if we frame the question differently, maybe that’s where the money comes

18. Closing Remarks

Shripad Tuljapurkur notes that it has been valuable to get some of the most talented members here raising tough questions, and sharing different perspectives and positions. He adds that there should be more of a policy focus in the field, particularly considering the tumultuous funding climate, and noting the linkage between the practical relevance of research and funding. “Innovation” has become the word of the day in the Bay Area, but a lot of academics are not innovating, they are just turning stuff into things people pay for. He ends with words of advice he tells his graduate students: it’s important to practice your elevator speech – you need to be able to tell people what you’re good for, and they’ll be happy to fund you.

Ken Wachter agrees that this has been one of the best meetings for the group, and notes that the meeting comes at a critical moment where there are a set of technical issues that need to be resolved, and a group here that is determined to resolve those issues. From the presentations during the conference, it is clear that the field of social science genomics is moving quickly, and that there are substantive studies that will generate wide scientific and public interest. The focus on genetics as it relates to the life course is breathing new life into the field, as it raises such questions as: where do advantages accumulate, what’s the nature of these causally remote phenotypes that are flowing through biology, and what’s the kind of causal story? These questions have been present for a long time, but we are now equipped with the tools to actually answer them; in the past, we have relied on and exhausted conventional correlation and regression analyses. The social science side of genomics is no longer a poor cousin to the biomedical side. We are conducting important research the biomedical community did not do and would not do. There is great potential in advancing this research, particularly given the strong network of collaboration observed here – almost every person in this room has had collaborative connections with at least one other person in the room.

Daniel Belsky expresses how great it is to gather at this conference and receive senior feedback. He remarks that the Social Science and Science Conference in Boulder, CO would be a good follow-up conference to attend.

Kathleen Mullan Harris wonders where the future of the field is headed, mentioning how methylation, telomeres, and gene expression were not touched heavily discussed, but all are exciting areas of interest, for which data is currently being gathered. Such data will inform how environmental changes bring about changes in genes and DNA.

David Steinsaltz notes that the progress that has been made in the field is at once impressive and intimidating. To create sensible ways to use genetic and social data together really puts increased pressure to those of us who are trying to get the theory straight. For a long time, studies could blame their results on insensible data; now we are headed in a direction where we cannot say that as much and need to devise the statistical tools to be able to determine and interpret results.
Appendix A: Program

Conference: Genes and Complex Human Phenotypes
May 26-27, 2016
UC Berkeley Campus, Social Science Matrix, Barrows Hall, 8th Floor
<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30</td>
<td>Shripad Tuljapurkar</td>
<td>Introduction</td>
</tr>
<tr>
<td>9:00</td>
<td>Daniel Belsky</td>
<td>Sociogenomic analysis of life attainments</td>
</tr>
<tr>
<td>9:30</td>
<td>Melinda Mills</td>
<td>Large-scale genomic meta-analysis identifies loci harbouring genes for human reproductive behaviour</td>
</tr>
<tr>
<td>10:00</td>
<td>COFFEE BREAK</td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Kathleen Mullan Harris</td>
<td>What Data Do We Need to Study Genes and Complex Human Phenotypes?</td>
</tr>
<tr>
<td>11:00</td>
<td>Discussion</td>
<td>Led by Mark Cullen: Implications of the research discussed at this conference for biomedical science</td>
</tr>
<tr>
<td>12:30</td>
<td>LUNCH</td>
<td></td>
</tr>
<tr>
<td>13:30</td>
<td>Ben Domingue</td>
<td>Mortality Selection in a Genetic Sample and Implications for Association Studies</td>
</tr>
<tr>
<td>14:00</td>
<td>Sid Krishna Kumar</td>
<td>GCTA And All That</td>
</tr>
<tr>
<td>14:30</td>
<td>David Steinsaltz</td>
<td>Random-effects estimates of heritability for simple and complex traits: Some statistical theory</td>
</tr>
<tr>
<td>15:00</td>
<td>COFFEE BREAK</td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td>David Rehkopf</td>
<td>Statistical learning approaches to identifying gene-environment interaction</td>
</tr>
<tr>
<td>16:00</td>
<td>Discussion</td>
<td>Led by Ken Wachter: &quot;The state of our understanding of the statistical properties of genome-wide analytic methods&quot;</td>
</tr>
<tr>
<td>17:30</td>
<td>Ends</td>
<td></td>
</tr>
</tbody>
</table>
Friday, May 27th

8:30    Dalton Conley    Do PCs Overcontrol Causal Effects?

9:00    Ken Wachter    Genetic Load and Polygenic Scores

9:30    COFFEE BREAK

10:00   Felix Tropf    Genetic Heterogeneity In Human Fertility And Educational
              Attainment By Country And Birth Cohort

10:30   Nicola Barban    Assortative Mating on Education: A Genetic Assessment,
              Nicola Barban, University of Oxford

11:00   Discussion    Led by Jeremy Freese: Implications of the research
              discussed at this conference for social science

12:30   LUNCH

14:00   Amal Harrati    Cognition, Retirement, and Genes

14:30   TBD

15:00   COFFEE BREAK

15:30   Discussion    Led by David Rehkopf: The Aging Process -- how do we
              measure it in this setting? How does it influence results
              derived from genetic studies, etc?

17:00   TBD

17:30   Ends
Appendix B: Meeting Participants

Marion Auoad
PhD Student, University of California, Berkeley

Lucia Aronica
Postdoc, Stanford University

Nicola Barban
Researcher, Oxford University

Daniel Belsky
Professor, Duke University

Eran Bendavid
Faculty, Stanford University

Boroko Bo
PhD Student, University of California, Berkeley

Carl Boe
Researcher, University of California, Berkeley

Gabriel Borges
PhD Student, University of California, Berkeley

Dalton Conley
Professor, New York University

Mark Cullen
Professor, Stanford University

Ben Domingue
Assistant Professor, Stanford University

Andrew Duong
Student, Stanford University
Ryan Edwards
Associate Professor, Queens College CUNY

Brian Finch
Research Professor, USC

Jeremy Freese
Researcher, Stanford University

Alison Gemmill
Student, University of California, Berkeley

Joshua Goldstein
Professor, University of California, Berkeley

Amal Harrati
Postdoc, Stanford University

Sid Krishna Kumar
Graduate Student, Stanford University

Ronald Lee
Professor, University of California, Berkeley

Bharathi Lingala
Statistical Programmer, Stanford

Melinda Mills
Professor, Oxford University

Sep Modrek
Instructor, Stanford University

Kathleen Mullan Harris
Professor, University of North Carolina, Chapel Hill

Maryam Nazemi
Engineer, PG&E

Xiaomin Niu
Visiting Researcher, Stanford University

John Openshaw
Instructor, Stanford University

Jinyuan Qi
Student, University of California, Berkeley
David Rehkopf  
Assistant Professor, Stanford University

Natalie Rusnak  
Student, University of California, Berkeley

Carolina Santamaria Ulloa  
Professor, University of Costa Rica

Ben Seligman  
PhD Student, Stanford University

David Steinsaltz  
Lecturer, Oxford University

Jacqueline Torres  
Scholar, University of California, San Francisco; University of California, Berkeley

Felix Tropf  
Researcher, Oxford University

Shripad Tuljapurkar  
Professor, Stanford University

Elizabeth Vasile  
Executive Director – CEDA, University of California, Berkeley

Ken Wachter  
Professor Emeritus, University of California, Berkeley

Julia Walsh  
Professor, University of California, Berkeley

Yohannes Woldeamanuel  
Senior Fellow, Stanford University

Mia Zhong  
Graduate Student, University of California, Berkeley