SOC10000: Introduction to Sociology

Sociogenomics

Wesley Wang
Purdue University

About Me!

Nice to meet everyone! My name is Wesley

- 1st year PhD student from Singapore
- Disciplinary Backgrounds
  - **English Linguistics**
    National University of Singapore
  - **Sociology & Demography**
    University of Oxford
  - **Sociology & Gerontology**
    Purdue University

Research interests: gene-environment interactions; stratification across the life course; causality

Class Activity 1

https://form.typeform.com/to/A4WqcgCA

Sociogenomics – what is it?

- The study of how genes influence social outcomes?
  - e.g., Educational Attainment
  - e.g., Age at first birth
  - e.g., Number of children ever born
- The study of how genes interact with social structures (e.g., social class; education systems) to produce differentiated outcomes?
  - Does one's genetic 'risk' for smoking weaken if you're from a wealthy (rather than deprived) family?
  - Does one's genetic 'ability' for education amplify if you're sorted into a resourceful/prestigious university?

Many other ongoing inquiries! Extremely rapid and evolving field

Sociogenomics – what is it **NOT**

- The study of how genes *determine* (rather than influence) social outcomes
- The scientific justification of why someone (or some social group) is "better/worse off" than the other
- The quest of identifying a set of 'best genes' to create the perfect human being

Although this has already been used for various nefarious purposes

- Defending criminal actions on the basis of genetic disposition ("I have no control over my actions")
- Eugenics: False view that human population can improve through selective breeding of genetically 'superior' individuals

The 3 Great 'Eras' of Sociogenomics

Sociogenomics is a rapidly evolving field, both theoretically and methodologically!

Sociogenomic methodological advances

1. Twin Studies
2. Candidate Gene Studies
3. Genome-wide Approaches

Twin Studies: Birth of Sociogenomics

Clarifications

- **Mono**zygotic (MZ) twins: Share on average 100% of genetic material (aka 'full' twins)
- **Di**zygotic (DZ) twins: Share on average 50% of genetic material (aka 'half' twins)
- Phenotype: Refers to the trait that we are measuring (e.g., education levels, income)
- Genotype: Refers to the 'genetic makeup' (loosely speaking) that we think relates to the phenotype of interest
- *Heritability: Proportion of variance in Phenotype explained by differences in Genotype

Twin Studies: Birth of Sociogenomics

Basic (but not wholly accurate) Intuition:

- Twin 1 has 100% genetic similarity. Twin 2 has 50% genetic similarity.
- If the *correlation* in income for Twin 1 is 0.5 and the correlation for Twin 2 is 0.2, then:

$$r_{diff} = 0.5 - 0.2 = 0.3$$

- Then this difference in correlation ($r_{diff} = 0.3$) *must* be due to this 50% difference in genetic similarity between twin 1 and 2

Twin Studies: Birth of Sociogenomics (cont...)

Rough intuition of how it works

- Assume that phenotype variance/variation (P) in a population can be broken down into three 'components'
  1. Additive genetic factors (A)
  2. Common environment shared by twins (C)
  3. Error induced by factors that are not A or C (E)

$$P = A + C + E \tag{1}$$

- **Goal**: Heritability (h^2) is the proportion of phenotype (P) *variance* explained by genetic differences (A)

$$h^2 = \frac{A}{P} \tag{2}$$

Twin Studies: Birth of Sociogenomics (cont...)

Rough intuition of how it works

- We want to know how much twins correlate with each other in their outcomes, given their shared genetic material
- For **DZ Twins**, we know they only share on average 0.5 of their genes (0.5A). They also share a common environment (C). Thus any correlation (how much Twin 1 and Twin 2 'share' similar outcomes) can be represented by:

$$r_{DZ} = 0.5A + C \tag{3}$$

In essence we are saying:
"Any shared outcomes (r_{DZ}) between twin 1 and 2 are due to their 50% shared genes (0.5A) and shared environment (C)"

Twin Studies: Birth of Sociogenomics (cont...)

Rough intuition of how it works

- For **MZ Twins**, we know they share on average 100% of their genes (A). They also share a common environment (C). Thus any correlation in outcomes can be represented by:

$$r_{MZ} = A + C \tag{4}$$

In essence we are saying:
"Any shared outcomes (r_{MZ}) between twin 1 and 2 are due to their 100% shared genes (A) and shared environment (C)"

Twin Studies: Birth of Sociogenomics (cont...)

- Let's recall all the information we've gotten so far

$$P = A + C + E \tag{1}$$

$$r_{DZ} = 0.5A + C \tag{3}$$

$$r_{MZ} = A + C \tag{4}$$

- Our objective is to figure the *heritability* (h^2)

$$h^2 = \frac{A}{P} \tag{2}$$

Twin Studies: Birth of Sociogenomics (cont...)

Solving for A, we can simply take the difference of equations (3) and (4), multiplied by 2:

$$A = 2(r_{MZ} - r_{DZ})$$

Thus, heritability can then be derived as follows:

$$h^2 = \frac{A}{P} = \frac{2(r_{MZ} - r_{DZ})}{P}$$

Respective values of r_{MZ}, r_{DZ}, and P can be obtained from the sample data used in analysis.

Twin Studies: Birth of Sociogenomics (cont...)

What does the heritability mean?

If $h^2 = 1.0$ (extremeeely unlikely):

- Interpretation: **ALL** variation in (say) education is *fully* attributed to genetic differences between people
- i.e., Genes *fully* determine your educational outcomes

If $h^2 = 0$ (also extremeeely unlikely):

- Interpretation: **NO** variation in (say) education can be explained in the slightest by genetic differences between people
- i.e., Genes predicts *nothing* about your educational outcomes

Twin studies assumptions

All methodological techniques have assumptions. What are some assumptions of Twin Studies?

1. Assumption of equal environments: both twins have the *exact* same experiences and upbringing
   - Highly unlikely – even within families twins are treated differently and experience different things
2. Independence of genes and environments: one's genetic makeup doesn't affect the environments he/she selects into
   - Highly unlikely – individuals with greater educational ability are likelier to select into schools with greater resources. Thus environments partially contain/reflect genetic effects
   - Overestimate A and underestimate C
3. Many more!

Class Activity 2!

https://form.typeform.com/to/nwkWS64k

Candidate Genes: A disgraceful past

Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits

John K. Hewitt

Received: 6 September 2011 / Accepted: 8 September 2011 / Published online: 18 September 2011
© Springer Science+Business Media, LLC 2011

The literature on candidate gene associations is full of reports that have not stood up to rigorous replication. This is the case both for straightforward main effects and for candidate gene-by-environment interactions (Duncan and Keller 2011). As a result, the psychiatric and behavior genetics literature has become confusing and it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge. The reasons for this are complex, but include the likelihood that effect sizes of individual polymorphisms are small, that studies have therefore been underpowered, and that multiple hypotheses and methods of analysis have been explored; these conditions will result in an unacceptably high proportion of false findings (Ioannidis 2005).

Because of this, the Editor and Editorial Board have increasingly erred on the side of caution in considering candidate gene association studies of complex traits. To avoid publishing findings that will not replicate, we recommend that authors conduct a direct replication analysis (Sullivan 2007), prior to publication, such that the same predictor(s), outcome variable, and statistical model are tested in an independent sample. Such replication does not guarantee that the result is correct as there are still many ways to obtain and replicate an artifactual result, but it does reduce the probability that the original finding was due to chance (a Type-I error) or to biases of other kinds that are more difficult to quantify. Direct replication should be a minimum requirement for candidate gene association

Candidate Genes: A disgraceful past

Establishing relationships between any 1 gene (e.g., APOE; MAOA) and complex traits

- Highly implausible that 1 gene can explain substantial amount of variation in (say) educational outcomes!
- Very few outcomes are actually the product of just 1 gene (e.g., Huntington disease)
- Therefore it would be more productive to examine the entire **genome** (i.e., the entire 'genetic makeup' of individuals)

Genome-wide Association Studies (GWAS)

Advances in technology has drastically reduced the cost of genome sequencing

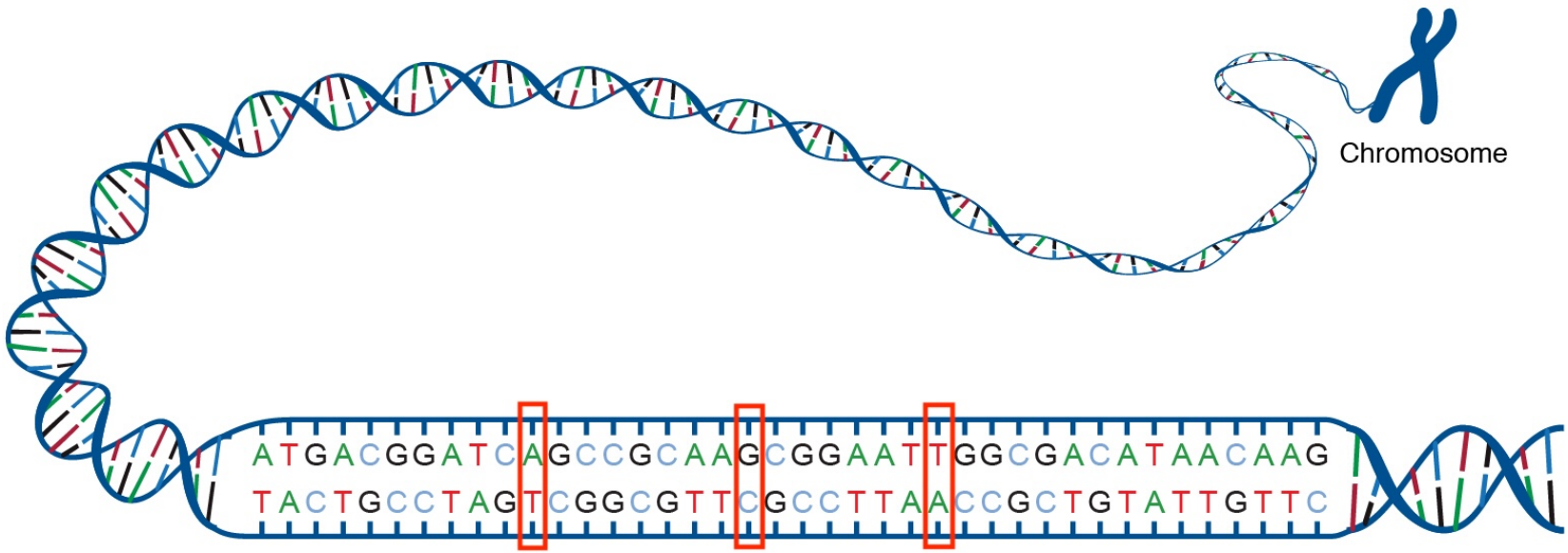
- Enables scientists to move beyond singular *genes*
- Focus is now to examine entire DNA, or more specifically, 'parts' of the DNA that differ between individuals
  - People are approximately 99.7% genetically similar
  - This means that technically we are only looking at ≈0.3% of the genome to account for differences between people

Genome-wide Association Studies (GWAS)

Alleles

Let's not dive too deep into biology – what you have to know:

- DNA comprised of 4 'bases' (or alleles)
  1. Adenine (A)
  2. Cytosine (C)
  3. Guanine (G)
  4. Thymine (T)
- Alleles usually come in *pairs* (i.e., a base pair)
  1. A-C
  2. A-A
  3. C-G
  4. etc...


Chromosome
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATTGTTC

Genome-wide Association Studies (GWAS)

Single nucleotide polymorphisms (SNPs)

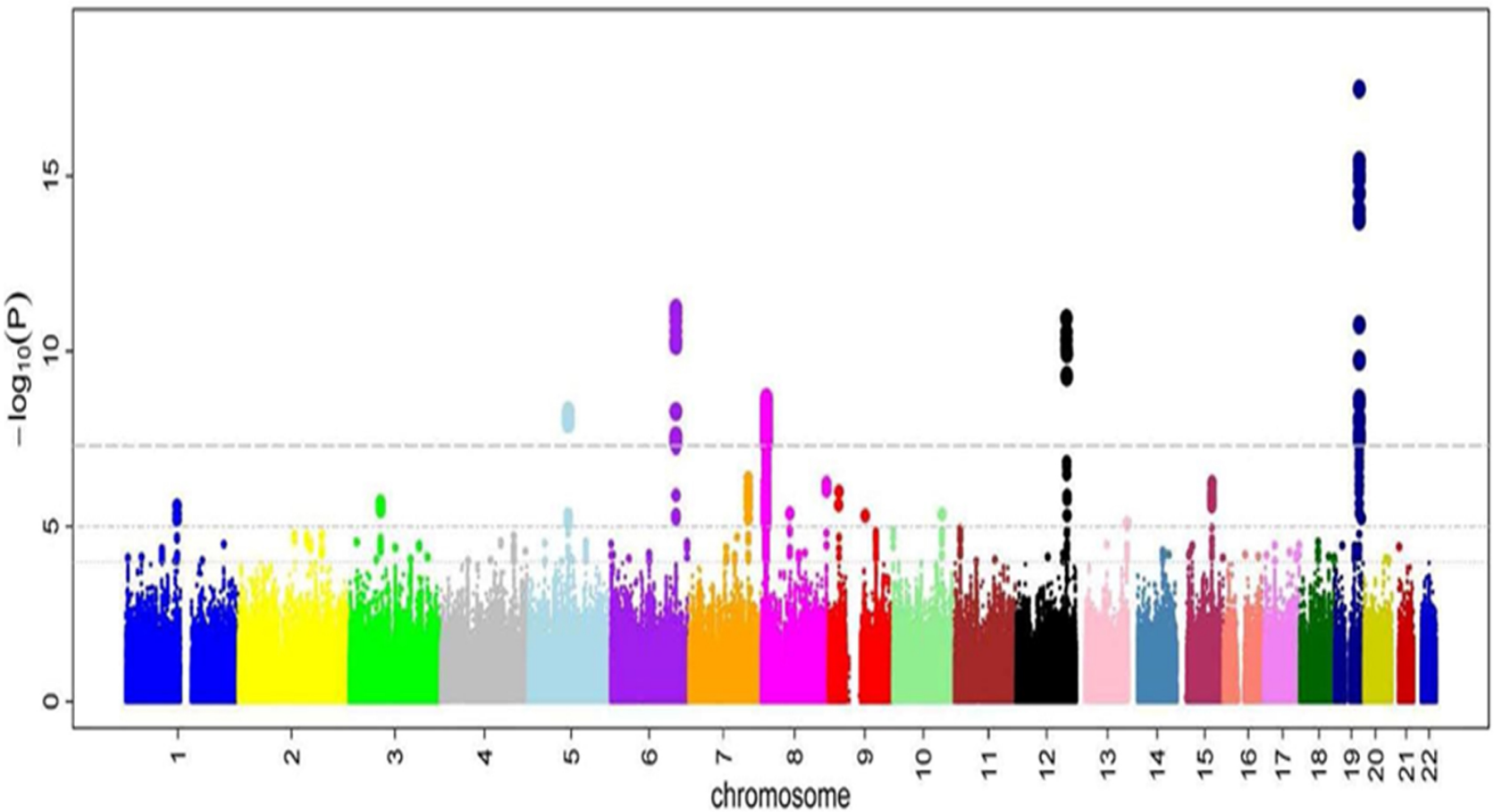
- Refers to specific *locations* (or loci) of our genome where alleles are known
  - e.g., The possible alleles at SNP rs6265 are A and G
  - e.g., The possible alleles at SNP rs7412 are T and C
- Each SNP comprised of a *base pair* which we know differs between people
  - Person 1 may have a A-T base pair
  - Person 2 may have a A-A base pair
  - Person 3 may have a T-T base pair
- Therefore at each SNP, people can have only 3 possible outcomes
  - Two alleles of interest (e.g., A-A)
  - One allele of interest (e.g., A-T)
  - No allele of interest (e.g., T-T)

Genome-wide Association Studies (GWAS)

Okay, so what exactly does a GWAS do...

- It 'scans' the whole genome to detect which SNPs actually matter for certain outcomes
- e.g., SNP number 123, and 456 contributes towards one's educational attainment, but not SNP number 789
- Till date, there are about 650 million known SNPs! So a very computationally intensive endeavour to conduct a GWAS

Manhattan Plots



Class Activity 3!

https://form.typeform.com/to/o9gic6qC

Genome-wide Association Studies (GWAS)

Okay, so what exactly does a GWAS do...

- It 'scans' the whole genome to detect which SNPs actually matter for certain outcomes
- e.g., SNP number 123, and 456 contributes towards one's educational attainment, but not SNP number 789
- Till date, there are about 650 million known SNPs! So a very computationally intensive endeavour to conduct a GWAS

Genome-wide Association Studies (GWAS)

Okay, so what exactly does a GWAS do...

- It 'scans' the whole genome to detect which SNPs actually matter for certain outcomes
- e.g., SNP number 123, and 456 contributes towards one's educational attainment, but not SNP number 789
- Till date, there are about 650 million known SNPs! So a very computationally intensive endeavour to conduct a GWAS

Why is this important? How can it be used in research?

Polygenic Risk Scores (PRS)

We can create a **polygenic risk score (PRS)** that tells us about a person's *genetic predisposition* towards a trait

- Having a high PRS score means that a person is *more* genetically likely to attain a certain trait
- Having a low PRS score means that a person is *less* genetically likely to attain a certain trait

Polygenic Risk Scores (PRS)

Rough (overly simplified!) intuition of how it's calculated

- Assumptions
  1. *Only* SNP 123 matters for educational attainment
  2. Having 1 allele A at SNP 123 is related to 0.3 additional years of education
- Comparing across persons:
  - Person A has 0 allele A at SNP 123
  - Person B has 1 allele A at SNP 123
  - Person C has 2 allele A at SNP 123

Person B is more likely to have $1 \times 0.3 = 0.3$ additional years of education compared to Person A

Person C is more likely to have $2 \times 0.3 = 0.6$ additional years of education compared to Person A

Polygenic Risk Scores (PRS)

- In reality, **many more SNPs** count towards an outcome
- Therefore repeat the above steps for ALL SNPs that were found to be significant in the GWAS

			Fred		Alice		Greg	
Genetic variant	*Effect allele*	*Effect size*	*Genotype*	*Effect*	*Genotype*	*Effect*	*Genotype*	*Effect*
rs12395	A	0.02	AA	+.02 (x2)	TT		AT	0.02
rs44346	G	-0.04	GT	-0.04	TT		TT	
rs72557	C	-0.05	CG	-0.05	CC	-.05 (x2)	GG	
rs18338	A	0.09	AT	0.09	TT		TT	
rs29849	T	0.004	TT	+.004 (x2)	CT	0.004	CT	0.004
rs43466	T	0.07	AA		TA	0.07	AA	
rs29457	G	-0.01	CC		CC		GC	-0.01
rs13458	C	0.015	AA		CA	0.0015	AA	
		Polygenic score:	0.043		-0.0245		0.014	

Important caveats

- GWAS and PRS scores are **not** portable across different ancestries (East Asians, Europeans, Africans etc)
  - Genetic makeup of Europeans and Africans are *drastically* different.
  - You can't calculate PRS scores for an African population using GWAS results for a European population and vice versa
- PRS scores are **not** deterministic
  - Having a high genetic score for smoking risk score does not guarantee that you WILL smoke.
  - Environments matter more than genetics for almost all cases

Important Caveats (cont...)

- GWAS (and therefore PRS) results are influenced by **social contexts**
  - Conducting a GWAS study in an East Asian population would most certainly find that certain SNPs are related to chopsticks use.
  - Does not mean necessarily that there are genes for chopstick use!
- Genetic scores often only explain only a small percentage of social outcomes
  - Okbay et al. [2022]: Education Attainment (12-16%)
  - Mills et al. [2021]: Age at First Birth (9–22%)
  - Akimova et al. [2025]: Occupational Status (7–10%)

Thank you for your attention!

Please feel very free to reach out anytime for any questions, comments, or coffee (preferably tea)!

Email: wang6429@purdue.edu

Github: https://github.com/wesleywj/SOC10000

References I

Akimova, E. T., Wolfram, T., Ding, X., Tropf, F. C., and Mills, M. C. (2025). Polygenic prediction of occupational status GWAS elucidates genetic and environmental interplay in intergenerational transmission, careers and health in UK Biobank. *Nature Human Behaviour*, 9(2):391–405.

Mills, M. C., Tropf, F. C., Brazel, D. M., van Zuydam, N., Vaez, A., eQTLGen Consortium, BIOS Consortium, Human Reproductive Behaviour Consortium, Pers, T. H., Snieder, H., Perry, J. R. B., Ong, K. K., den Hoed, M., Barban, N., and Day, F. R. (2021). Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour. *Nature Human Behaviour*, 5(12):1717–1730.

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D. A., Ahlskog, R., Magnusson, P. K. E., Oskarsson, S., Hayward, C., Campbell, A., Porteous, D. J., Freese, J., Herd, P., Watson, C., Jala, J., Conley, D., Koellinger, P. D., Johannesson, M., Laibson, D., Meyer, M. N., Lee, J. J., Kong, A., Yengo, L., Cesarini, D., Turley, P., Visscher, P. M., Beauchamp, J. P., Benjamin, D. J., and Young, A. I. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, 54(4):437–449.