SISG 2022: Module 11
Session 9: Bioethics and legal issues

1. The NIH awards research grants using tax-payer money. Because the public is essentially paying for the research, the NIH has stipulated that all data collected in these funded studies must be made available to other researchers and stored in a communal database, including genomic data. An Indigenous community has proposed a compelling genetic epidemiology research study, but does not want the genetic data deposited in the database for other researchers to access. The concern is that genetic data identifies specific individuals, that the community is identifiable, and that researchers accessing the genetic data may use it for research studies that the community does not approve and does not agree with.

Bioethical category	Considerations
Beneficence	Collecting genetic data can help lead to better understanding of biology of disease and lead to therapies that can apply to this indigenous community. It can also help built relationships between institutions and indigenous communities.
Non-maleficence	If the data is shared without permission with other parties, this can erode trust. It's also possible that harm can be caused by restricting the opportunities for research in a community or forcing them to adhere to standards that were developed for and by non-Indigenous people.
Autonomy	Autonomy is important to consider when thinking about who has the power to give access to data. There is a possibility that Indigenous communities only want their data to be used in a certain way and they should have more power to decide who else can use it or benefit from it.
Justice	Justice is about treating people/groups equitably, not equally. In this case, making an exception for Indigenous communities may be more equitable.

   a. Use the table to map bioethical considerations for whether the NIH should still award this grant even if the genetic data are not deposited in the database. Consider at least two different stakeholder viewpoints.

   b. What other options could there be besides simply funding or not funding the research study?

      There can be more opportunities for input from the community – investigators may need a letter of support from the community or a member should be on the council for grant decisions regarding research using the data. NIH could even start a new initiative devoted to indigenous health and genomic information, with data sharing agreements specific to this community.

SISG 2022: Module 11
Session 10: Rare variation

1) You just got a large grant to identify rare variants associated with type 2 diabetes. You have colleagues around the world that can give you access to DNA from their case-control studies. If you were to design a study to identify rare (allele frequency <1%) variants associated with type 2 diabetes, what approach would you take and why?

    a. High-depth whole genome sequencing
    b. Low-depth whole genome sequencing
    c. Whole exome sequencing
    d. GWAS chip and imputation
    e. Exome chip (custom array)

There are pros and cons to different methods, as seen below. One viable option is combining methods, like doing a subset with HD-WGS, and then targeted sequencing for others or HD-WGS and then GWAS

	Advantage	Disadvantage
High-depth WGS	can identify nearly all variants in the genome with high confidence	very expensive
Low-depth WGS	cost-effective and useful approach for association mapping	has limited accuracy for rare-variant identification and genotype calling; compared to deep sequencing, is subject to power loss if the same number of subjects is sequenced
Whole-exome sequencing	can identify all exonic variants; is less expensive than WGS	is limited to the exome
GWAS chip and imputation	inexpensive	has lower accuracy for imputed rare variants Will miss any variants unique to your sample
Exome chip (custom array)	much cheaper than exome sequencing	provides limited coverage for very rare variants and for non-Europeans is limited to target regions

SISG 2022: Module 11
Session 11: Gene x Environment Interactions

1) You are conducting a GxE interaction study, where the environmental exposure is smoking. Your colleagues have shared their data with you, which means you can include 25,050 subjects in your study! You need to harmonize the smoking variable across studies. The studies, their sample size and study-specific questions related to smoking can be found in the table. You are trying to build the biggest dataset you can, but you must be able to use the same definition of smoking. What are the samples sizes you could have in your study if you used the following definitions for your "smoking" exposure?

    a. Cigarettes per day - – 2,500 (can potentially add Study 2 and 7: 4,500)
    b. Ever smoker 10,500 (Studies 2, 3 and 7)
    c. Current smoker - 16,660 (Studies 1, 2, 4, 5, 6, 7)

(a)

Study (*N*)	Smoking-related questions	Possible responses
Study 1 (2,500)	1. Do you currently smoke cigarettes?	Y/N
	2. If yes, how many cigarettes per day?	###
Study 2 (1,200)	1. Have you smoked more than 100 cigarettes in your lifetime?	Y/N
	2. If yes, do you currently smoke?	Y/N
	3. If yes, how many packs per day do you smoke?	###
Study 3 (8,500)	1. Have you ever smoked?	Y/N
Study 4 (1,250)	1. Do you currently smoke?	Y/N
Study 5 (4,200)	1. Do you smoke?	Y/N
	2. When did you first start smoking regularly?	Past year; 1–5 years ago; >5 years ago
Study 6 (6,600)	1. Have you smoked tobacco in the past month?	Y/N
Study 7 (800)	1. Have you ever smoked regularly?	Y/N
	2. If yes, do you still smoke?	Y/N
	3. If yes, how much do you smoke a day?	1–10 cigarettes, 11–20 cigarettes, 21–30 cigarettes, >30 cigarettes

SISG 2022: Module 11
Session 12: Risk Prediction and Population Screening

1) Why would a polygenic risk score developed in a European ancestry cohort be unreliable for a person who does not have recent European ancestors? Hint: think about the mechanics of GWAS that give specific SNPs and loading values.

   Differences in LD patterns, allele frequency differences, different effect sizes in different populations.

2) Discuss the ethical and social implications of using polygenic risk scores for embryo selection
   a. How should OrchidHealth handle rapid scientific developments? What happens if after an embryo is selected, new research comes out that shows that high PRS for one disease is inversely related to another disease?

   Many considerations including disparities in accuracy, inability to accurately predict complex diseases, access ($$) issues, unfinished genetic discoveries

---
Population Screening
---

3) Determine how whole population screening compares to cascade screening for cost-effectiveness. In the general US population, the collective variants causing Familial Hypercholesterolemia (FH) are found at a frequency of 1/250 (0.004). All first-degree relatives of a proband (patient) with FH have a 0.50 frequency of also having FH. Assume each genetic test costs $250. Genotyping errors (leading to false positive test result) occur at a rate of 0.1%.
   a. How much does it cost to detect one person with FH in the general population compared to among first degree relatives of a proband?
      250 tests to discover one , $250 x 250 = $62,500
      A parent with FH has a 50% chance of passing on the allele with FH allele. $500 to detect .

   b. Consider the error rate of genotyping for this platform. How many false positives do you expect per true positives in the entire population compared to in cascade screening?
      For every 999 True positives, there is a false positive, based on an error rate of 0.1%, so for 250 tests (1 positive in the population), there will be .25 false positives. In familial cascade screening, for 1 true positive, there will be 0.002 false positives.

   c. Variants in three genes are responsible for 80% of FH cases. What are strategies for identifying FH cases without variants in these three genes?

      Familial cascade screening of cases without these variants. Genotype a few cases. Then try to test family members of those who are cases but don't have mutations in one of those three genes.

SISG 2022: Module 11
Session 13: Mendelian randomization

1) Explore MR-Base (http://www.mrbase.org) to conduct your own MR study.

2) Run an MR study of body mass index and lung cancer risk following the example in class.

MR assumptions hold for the highlighted scenarios (A-C):

> In which examples (a-f) below do the MR assumptions not hold for assessing the association between exposure (X_1) and outcome (Y)? Why? Why not?

a. genetic variants → exposure → mediator → outcome

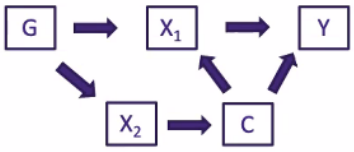
b. genetic variants → mediator → exposure → outcome

c. genetic variants → exposure → outcome
exposure ↓ related variable

d. genetic variants → related variable → outcome
related variable ↓ exposure

e. genetic variants → outcome → exposure

f.



EPIDEMIOLO