SISG 2021: Module 10

Session 1: Introduction

Each person, please introduce yourself to the other members in your group:

1. Your name and pronouns. Your position (student, researcher) and affiliation (what University or institute).
2. What are your strengths in your training so far? (i.e., is your background in genetics, biostatistics, law?)
3. What prompted you to take this course? What are you hoping to learn?

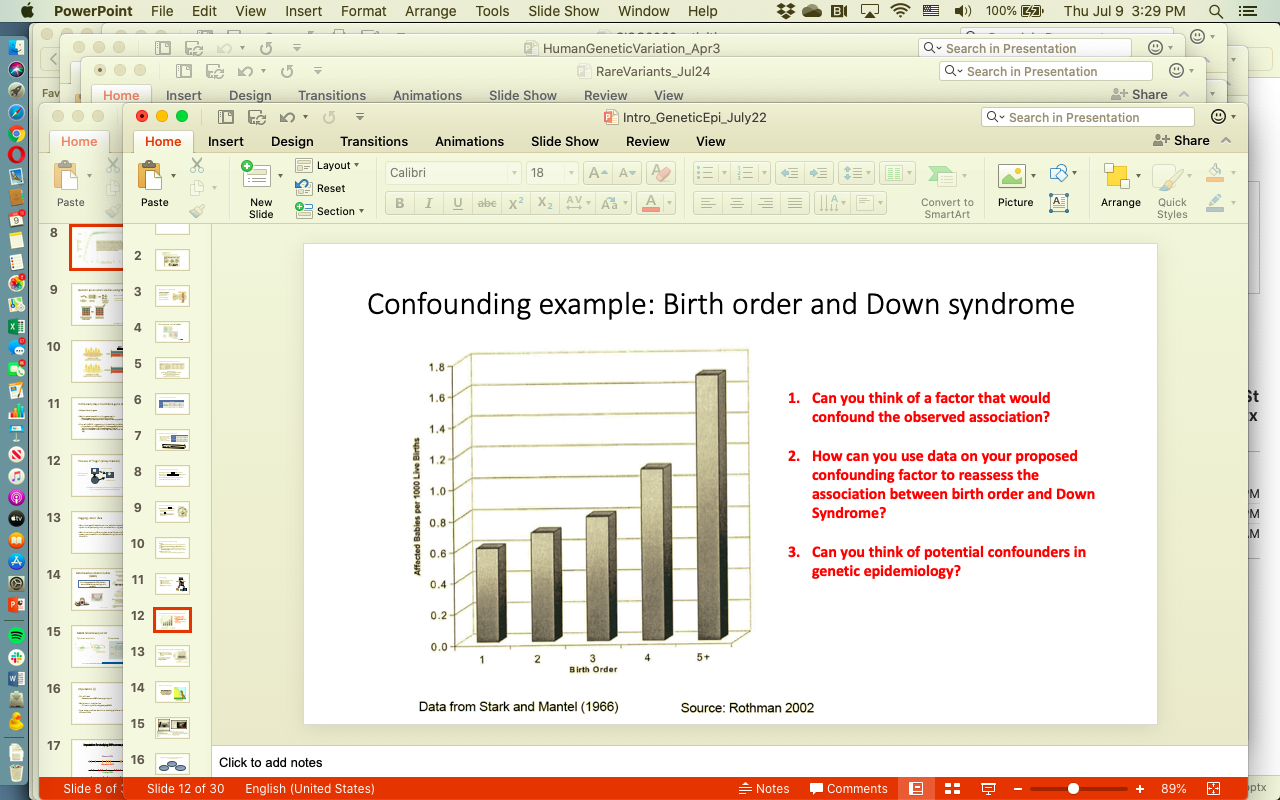
Once everyone is introduced, discuss in your group:

1. Why do we study the role of genetics in human disease?

SISG 2021: Module 10

Session 2: Introduction to Genetic Epidemiology

The following questions are based on this figure:



1. One of the most significant factors confounding the observed association is the maternal age. It is associated with both the exposure (Birth order) and the outcome (Incidence of Down Syndrome)

2. One approach is to stratify the study population based on the maternal age and investigate the association of interest within each stratum.

3. When studying the genotype-phenotype association in genetic epidemiology research, the most significant confounder is population structure, which represents the difference in allele frequency between the subpopulations.

SISG 2021: Module 10

Session3: Human Genetic Variation

1. A recent study sequenced the genome of 2,504 individuals and identified 84.7 million SNPs (single nucleotide polymorphisms) between the participants. On average, each individual carried 3.5-4.3 million SNPs each. About 0.5% of those SNPs were in coding regions of genes. Remember, 1.5% of the genome is in a coding region. Why might only 0.5% of variants be in coding regions compared to what would be expected if SNPs were randomly allocated throughout the genome?

Heterogeneity of SNPs distribution across the genome suggests the coding regions are more conservative. The polymorphisms in coding region may lead to severe functional consequences (missense mutations, nonsense mutations, frameshift mutations), and may be eliminated before they get the chance to be passed to the next generation.

1. Match the genetic term with the definition:
2. Nonsense
3. Heterozygous
4. Exon
5. Allele
6. Synonymous
7. Missense
8. Non-coding region
9. Haplotype
10. Autosomal
11. Phenotype
12. Genotype
13. Frameshift
14. Intron
15. Homozygous

d\_\_\_ Alternative forms of a gene or DNA base.

k\_\_\_ Genetic makeup of an individual at a particular DNA location based on both alleles.

b\_\_\_ Genotype consisting of two different alleles at a particular location.

e\_\_\_ DNA base change that does not change the translated amino acid.

n\_\_\_ Genotype consisting of two of the same alleles at a particular location.

j\_\_\_ Observable characteristics resulting from a genotype.

i\_\_\_ Concerning the 22 pairs of chromosomes that are not sex chromosomes.

m\_\_\_ Portion of gene that does not code for amino acids and appears in between exons.

l\_\_\_ Insertion or deletion mutation that changes the whole subsequent sequence of amino acids by changing the 3-codon groups for generating amino acids.

c\_\_\_ Portion of gene that encodes amino acids.

g\_\_\_ Section of DNA that does not become protein.

a\_\_\_ Substitution of a single DNA base that causes a stop in protein production.

f\_\_\_ DNA base change that changes the translated amino acid.

h\_\_\_ Set of DNA variations at several positions that are inherited together.

1. Look up “rs7412” in dbSNP (https://www.ncbi.nlm.nih.gov/snp/). <https://www.ncbi.nlm.nih.gov/snp/rs7412>
   1. What are the DNA bases identified at this location? C/T
   2. What gene is this SNP located in? *APOE*
   3. What is the effect of this SNP on the amino acid sequence? Missense Mutation
   4. Click on the “frequency” tab to the left. What is the frequency of the minor allele (less common allele) in the 1000 Genomes study overall? How do these frequencies differ by ancestral subgroup within this study? Total: 92.5% C/7.5% T

SISG 2021: Module 10

Session 4: Hardy-Weinberg Equilibrium and Linkage Disequilibrium

1. Sickle Cell Anemia is characterized by fatigue, pain, arthritis, frequent bacterial infections, and sudden pooling of blood in the internal organs that can lead to tissue damage. It is caused by a SNP in the hemoglobin gene that causes red blood cells to form sickle shapes instead of round, donut shapes. The SNP is a missense mutation (T>A) that replaces a glutamine with a valine and causes hemoglobin molecules to clump.
2. You are conducting a study and find that the A allele has a frequency of 20% among adults. Use Hardy-Weinberg Equilibrium equations to calculate how many people out of 1,000 you would expect to have each genotype. Remember the HWE equations are

Assuming 20%A, 80% T:

Based on HWE: AA: 0.22 = 0.04; TT = 0.82 = 0.64; AT = 2x0.2x0.8 = 0.32

Estimated Genotype frequency: AA: 40; TT: 640; AT: 320.

1. In the study of these same 1,000 people, you find that actually 605 people have the TT genotype, 390 have the TA genotype, and 5 people have the AA genotype. Is this a statistically significant deviation from what you would expect based on the allele frequencies? We calculate the chi-square value with the following equation and compare it to a chi-square distribution with 1 degree of freedom (3.841).

Estimated frequency: AA: 40; AT: 320; TT: 640

Observed frequency: AA: 5; AT: 390; TT: 605

X2 = [(5-40)2/40 + (390-320)2/320 + (605-640)2/640] = 47.8516

It is a statistically significant deviation from the HWE

1. What might be happening in the population to give you this HWE pattern?

Sickle Cell Anemia is common in the same global locations where malaria has historically been very common (Africa, India, Middle East). The malaria parasite cannot survive in red blood cells that sickle. In fact, in these regions, children with the TA genotype are most likely to survive to adulthood. This is called the “heterozygote advantage.” Children with the TT genotype are more likely to die from malaria, children with the AA genotype suffer from sickle cell anemia, and children with the TA genotype have natural defenses against malaria because their cells sickle under pressure (infection) but remain round when not stressed.

1. Why do we only look for LD between SNPs that are on the same chromosome?

Chromosomes are segregated independently during meiosis, so SNPs on different chromosomes are not able to be physically linked.

1. Here we will explore LD using the NCI LDLink online tools. You can find this website at ldlink.nci.nih.gov/?tab=home. We have a lot of different tools to explore, but here we will use the LDpop tab.
   1. Compare LD for the two SNPs that define alleles in two important genes affecting drug metabolism. These are rs776746 and rs2740574. Type these into the two SNP boxes (variant RS number) and select “(ALL) all populations”, “R2”, then “calculate.” After a few seconds, you should see a map of the world with tear drops showing different populations that have been studied, each labeled by the population. What is the LD R2 value among the British in England and Scotland (GBR) compared to the LD R2 value among the Luhya in Webue, Kenya (LWK) and compared to among Colombians from Medellin, Colombia (CLM)? You can find the details for each population in the table, and by clicking on the corresponding tear drop.

R2 among GBR: 0.5864

LWK: 0.0276

CLM: 0.225

* 1. Why might these LD values be so different between these populations?

Because LD pattern can reflect the evolutionary history of a population and may be affected by various factors, including nature selection, genetic drifting, inbreeding, non-random mating, bottleneck effect, and founders’ effect.

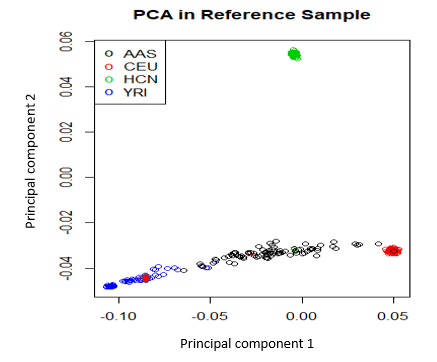
* 1. Most genetic studies occur in participants of European descent. When we are studying drug metabolism affected by these two SNPs, why might knowledge from these studies have limited utility across ancestral populations?

The research results based on European-ancestry population may not be applicable to other populations, as the linkage between those two SNPs in other populations of interest is medium to low.

SISG 2021: Module 10

Session 5: Population substructure

1. Below is a plot of principal component 1 vs principal component 2 in a sample of people from 4 populations: African Americans from the Southeast (AAS), Europeans from Utah (CEU), Yorubans from Nigeria (YRI), and Han Chinese from Beijing (HCN). Each dot represents one person and each person is color-coded based on their self-described race.



1. What populations are separated by principal component 1? What populations by principal component 2?

Most individuals have a negative PC2 value, but HCN population has positive PC2; AAS, CEU and YRI populations can be separated using PC1, although there are some potential misclassifications.

1. Why do we see tight-ish clusters of the three corner populations (blue, red, green), but the black circles are spread across the axis from blue to red along principal component 1?

AAS population is internally heterogeneous. Probably they are nearly admixed population that has features between YRI and CEU.

1. Notice the red dot in the lower left corner among the blue dots. What might be happening here? Remember what color refers to compared to what the principal components measure.

Potential misclassification of self-described race/Distinctive results of self-described and genetically determined race.

1. Where on this plot might you see people who describe their ancestry as Chinese American (ancestors from both European and Chinese populations)?

Somewhere between the red and green cluster.

1. What are pros/cons of using self-described race vs genetic ancestry in epidemiology studies? Think of what each can tell you based on the questions you are trying to ask.

Self-described race, pros: Easily obtained via interview/questionnaire; may reflect the character of race/ethnicity beyond genetic background (culture, shared lifestyle); cons: Potential misclassification as self-determined race may be distorted.

SISG 2021: Module 10

Session 6: Study design and imputation

1. Explore the breakdown of genetic ancestry in GWAS as reported on the website [https://gwasdiversitymonitor.com](https://gwasdiversitymonitor.com/). What do you notice about recent trends? What populations seem over- and under-represented in genetic studies?

Majority of the genetic association studies were based on European ancestry population. In recent years, although number of studies in Asian and African American population increased, the minority populations (Africans, African Americans, Asians, and Hispanics) are still under-represented.

1. What are your ideas for how we can we increase the diversity of study participants in genetic epidemiology?

SISG 2021: Module 10

Session 7: Association studies calculations and interpretations

1. You conduct a case/control study among 1,656 participants. You are particularly interested in the odds ratio for the outcome among those homozygous for the C allele vs. either heterozygous or homozygous for the T allele. You genotype everyone for that particular SNP and find the following genotype frequencies among your cases and controls.
   1. Calculate the odds ratio and 95% confidence interval for the odds of having the outcome among CC vs TT/TC genotypes.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cases** | **Controls** | **Total** |
| **TT+TC** | 158 | 392 | 550 |
| **CC** | 20 | 86 | 106 |
| **Total** | 178 | 478 | 1,656 |

Remember the following equations:

OR

s.e(log(OR))=

Lower limit of 95% confidence interval:

Upper limit of 95% confidence interval:

Odds ratio = (158x86)/(392x20) = 1.73.

Log(OR) = 0.548

\*Note: this is the natural log (ln). ln R: log() = ln(), use log10() to get log base 10 in R

SE(log(OR)) = (1/158 + 1/392 + 1/20 + 1/86)^1/2 = 0.2655

Upper limit of the CI: exp(0.5481+ 1.96 x 0.2655) = 2.91

Lower limit of the CI: exp(0.5481- 1.96 x 0.2655) = 1.03

* 1. Turn this result into a sentence describing the association between the CC genotype and odds of the outcome compared to the TT/TC genotypes.

Relative to the population with CC genotype at the locus, those who with TT or TC genotype would have 1.73-fold of the odds to develop the outcome.

1. You conduct a case/control study using an additive inheritance model. Your logistic regression output is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients |  |  |  |  |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 7.75 | 1.493 | 34.480 | <2e-16\*\*\* |
| genotypeAdd | 1.504 | 0.251 | 6.714 | <2e-16\*\*\* |

* 1. Determine the odds ratio for the odds of the outcome among participants with 2 copies of the allele of interest (genotypeAdd =2) compared to the odds among participants with 1 copy of the allele of interest (genotypeAdd = 1).

Odds ratio = exp(1.504) = 4.50

* 1. Use the std.error to determine the 95% confidence interval for that odds ratio estimate using the following equation with the standard error (s.e.):

Lower limit of 95% confidence interval:

Upper limit of 95% confidence interval:

Lower limit = exp(1.504 - 1.96 x 0.251) = 2.75

Upper limit = exp(1.504 + 1.96 x 0.251) = 7.36

* 1. Bonus: Determine the odds ratio for the odds of the outcome among participants with 2 copies of the allele of interest (genotypeAdd =2) compared to the odds among participants with no copies of the allele of interest (genotypeAdd = 0).

OR = exp(2 x 1.504) = 20.25

1. You conduct a quantitative association study of bone mineral density using an additive genotype model. Your linear regression output is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients |  |  |  |  |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 7.75 | 1.493 | 34.480 | <2e-16\*\*\* |
| genotypeAdd | 1.504 | 0.251 | 6.714 | <2e-16\*\*\* |

* 1. What is the average change in bone mineral density for every additional allele of interest?

REMEMBER WE ARE USING A LINEAR REGRESSION HERE.

For every additional allele of interest, the bone mineral density increases by 1.504 units

* 1. Among people homozygous for the allele of interest (genotypeAdd=2), what is the average bone mineral density?

Average bone mineral density among rare homozygous = 7.75 + 1.504 x 2 = 10.758

SISG 2020: Module 10

Session 8: Genome wide association studies

1. Explore the NHGRI-EBI GWAS catalog: (<https://www.ebi.ac.uk/gwas/home>). This website will introduce you to existing genome wide studies on many different phenotypes.
2. Using the GWAS catalog, determine what the SNP rs6025 has been associated with in previous studies.

Five trait(s) have been associated with SNP rs6025 based on the previous GWASs: venous thromboembolism; Antithrombotic agent use measurement; abnormal thrombosis, deep vein thrombosis, Ischemic stroke, pulmonary embolism, stroke, venous thromboembolism; inflammatory bowel disease; peripheral arterial disease

1. Explore the Global Biobank Engine ([https://biobankengine.stanford.edu](https://biobankengine.stanford.edu/)), which has collated GWAS results on a wide range of phenotypes based on large biobanks (UK Biobank, Biobank Japan, Million Veterans Program). Using this resource, what associations do you see with rs6025?

TTE phlebitis & other coagulation disorders

DVT & blood clots

PheWAS:

Timeline

Description automatically generated

Clicking on the top result shows GWAS results filtered by low p-value:

Chart

Description automatically generated

SISG 2021: Module 10

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 |  |
| Non-maleficence |  |
| Autonomy |  |
| Justice |  |

* 1. 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.
  2. What other options could there be besides simply funding or not funding the research study?

Many possible answers. Discussion rooms

SISG 2021: Module 10

Session 10: Rare variation

1. You have a large, but not unlimited budget. You have colleagues around the world that can give you access to DNA from their breast cancer case/control studies. If you were to design a study to identify rare (allele frequency <1%) variants associated with breast cancer, what are the advantages and disadvantages of each approach? What approach would you take and why?
   1. High-depth whole genome sequencing

Pros: Accurate detection of rare variants; this is a well-studies phenotype, so it is worth the cost to look at new variants, or maybe you just sequence participants belonging to understudied populations. Good option if you want to perform burden tests, SKAT tests, etc.

Cons: Cost!

* 1. Low-depth whole genome sequencing

Pros: More affordable than high-depth WGS and has many of the same advantages.

Cons: Less precision for rare variants (harder to distinguish heterozygosity at a site from a sequencing error)

* 1. Whole exome sequencing

Pros: Generally more affordable than WGS. Hits in exons tend to be more interpretable.

Cons: Can’t capture all parts of some exons. Most GWAS hits aren’t in exons,

* 1. GWAS chip and imputation

Pros: Your colleagues probably already have the GWAS chip data, so you only need to pay for the imputation.

Cons: Imputation results in loss of power, especially for rare variants

* 1. Exome chip (custom array)

Pros: Can focus on candidate SNPs or genes

Cons: Candidate studies have resulted in a lot of spurious associations / non-reproducible results. Custom arrays are usually built around data from European populations, which have already been well studied.

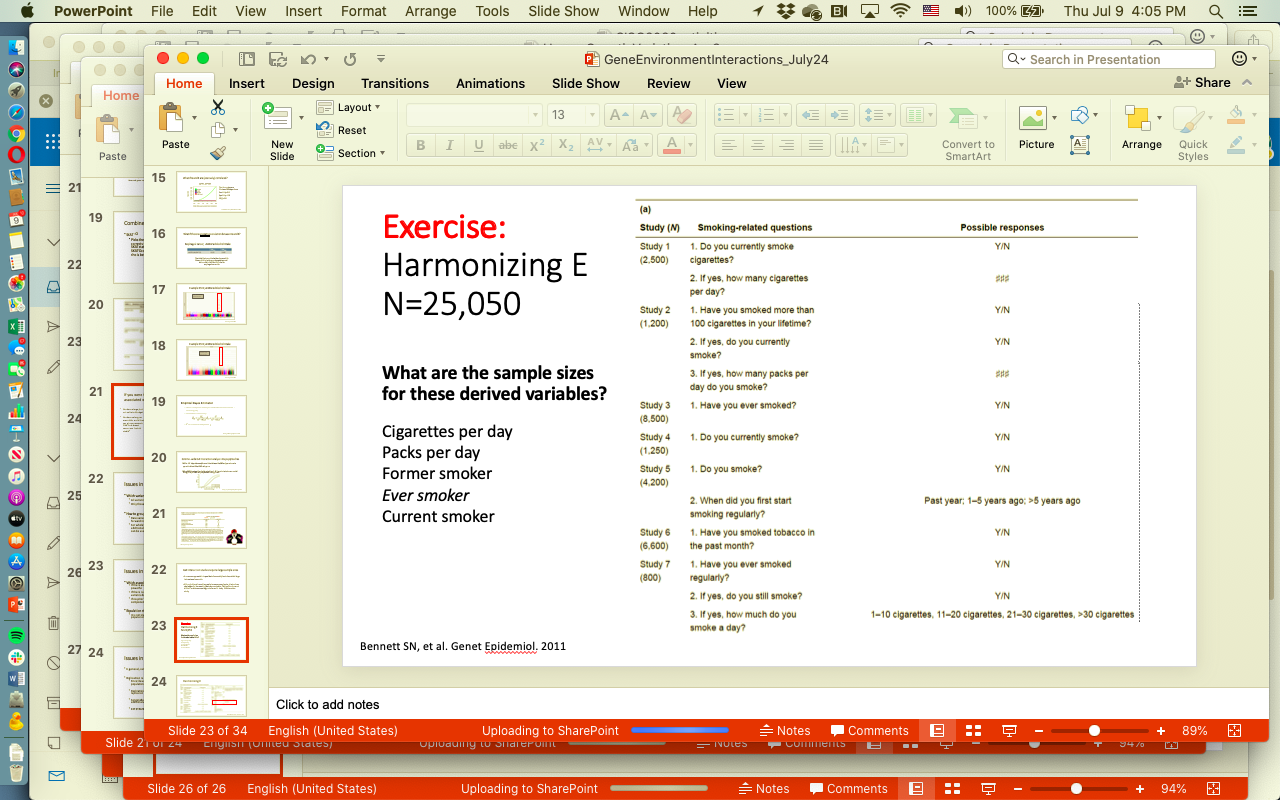
Text, application

Description automatically generated

SISG 2021: Module 10

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?
   1. Cigarettes per day – 4,500 (Study 1 and 7 and convert 2)
   2. Ever smoker – 10,500 (Studies 2, 3 and 7 – assuming people who currently smoke aren’t automatically considered “ever smokers”)
   3. Current smoker – 16,660 (Studies 1, 2, 4, 5, 6, 7)



SISG 2021: Module 10

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.

Some susceptible loci of a disease may only exist in a specific population. Also, the effect size of the SNPs on disease risk may vary by different ancestry. So the polygenic risk score constructed based on European ancestry population may not accurately predict the disease risk in a diverse population. If tag SNPs are used to develop the score, they may have lower LD with effect SNPs in other populations and, therefore, have less power to predict the outcome.

1. Determine how population allele frequencies affect implementation decisions for population screening: Cystic fibrosis causes thick and sticky mucus that affects the lungs, digestive system, and reproductive system. Because of this, people with cystic fibrosis have shorter-than-normal life spans. Early initiation of treatment may slow disease progression. Your colleagues are proposing screening the all infants born in the US for cystic fibrosis using genetic sequencing. Having two variants in the CFTR gene leads to cystic fibrosis (recessive inheritance).
   1. In the general US population, variants in the CFTR gene are found at an overall frequency of 0.017. How many infants in 100,000 would you expect to be able to identify with cystic fibrosis in order to initiate early treatment?

100,000 \* 0.0172 = 28.9

* 1. Variants in CFTR are much more common in some global populations than in other. For example, variants in the CFTR gene are found at a frequency of 0.204 in the Ashkenazi Jewish population. In a population of 100,000 infants of Ashkenazi Jewish ancestry, how many would you expect to have cystic fibrosis?

100,000 \* 0.2042 = 4161.6

* 1. Consider that each sequencing test costs $1000, what recommendations might you make for screening all infants in the US or all infants with Ashkenazi Jewish ancestry? Use the bioethical framework to help you put the economics in context of beneficence, non-maleficence, autonomy and justice.

Since DNA sequencing is expensive, we should consider sequencing the infants in the population with relatively higher risk (Ashkenazi Jewish ancestry). We should also consider sequence the infants from other families that have a family history of the disease, as they have a higher inherited risk.

* 1. How might your assessment change for using genotyping chips instead of gene sequencing? Remember, chips will only look at a few pre-specific variants in CFTR (10 most common variants identified in European populations), whereas sequencing will give you data on every DNA base position in the gene. The genotyping chip only costs $50/infant.

Genotype chip is much cheaper than sequencing technique. As hundreds of thousands of SNPs can be tested using one genotype chip, we can design one chip for testing multiple rare diseases, including cystic fibrosis. This approach can be used to screen all infants in the US, as it can lead to an early detection of the disease with a lower cost.

SISG 2021: Module 10

Session 13: Mendelian randomization

1. Explore MR-Base ([http://www.mrbase.org](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.

Example output. This shows that the overall effect is not significant but you might see something different based on the studies you used. (I used a study of sibling sets, which is not ideal)

Graphical user interface, application

Description automatically generated