



PERSONALGENOMES.ORG™

Personal
Genome
Project



Personal Genome Project Study Material

This study material contains six modules. Each of these modules contains the same study guide that is presented within the PGP Enrollment Exam. The questions on the enrollment exam are not included here.

Some of the material in this study guide is written assuming that the reader is currently taking the PGP enrollment exam. For information, please visit: www.personalgenomes.org

The modules are as follows:

Module 1: Human Subjects Research and Informed Consent

Module 2: Genetic Concepts

Module 3: PGP Enrollment Procedures

Module 4: PGP Participation, Sample Collection & Data Release Procedures

Module 5: Potential Consequences of Receiving Personal Genome Data

Module 6: Risks Associated with Making Your Data Public

Module 1:

Human Subjects Research and Informed Consent

Introduction

This first module is a brief review of human subjects research and informed consent. These topics are not unique to the Personal Genome Project, but they explain why taking this enrollment exam is an important part of your participation in this project. To allow us to share genetic and health data publicly, the PGP has pioneered a new method in human subjects research called “open consent”. This consent process places a high importance on your ability to understand and consent to the risks involved in making your data and samples public.

Human subjects research and the Belmont Report

Human subjects research has been a critical aspect of scientific advances in human health -- for example, drugs and other therapies must go through several rounds of testing in humans before they can be used by doctors to treat a disease.

Unfortunately, there have been cases where researchers have abused their relationship with research subjects. One infamous example of this is the Tuskegee Syphilis Study (1932-1972), where researchers studied cases of syphilis afflicting poor black men in Tuskegee, Alabama. These men believed they were receiving free health care from the US government -- in actuality, the researchers never treated them with penicillin (which by 1947 had become the standard treatment for syphilis) and were never informed of the disease.

In 1974 the United States Congress responded to these issues by establishing the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. In 1979 this commission published the Belmont Report, which guides much of current human subjects research. The Belmont Report outlines three fundamental ethical principles which should guide all human subjects research: respect for persons, beneficence, and justice.

1. Respect for persons:

Individuals should be treated as autonomous agents, and persons with diminished autonomy are entitled to special protection. For individuals capable of deliberation and self-determination, this means they must be given adequate information regarding the study and that their considered opinions and choices should be respected. Some individuals are not capable of making autonomous decisions (due to youth, mental impairment, or vulnerability to exploitation) -- research studies must treat such individuals with additional protection, possibly excluding them from activities which may harm them.

The Personal Genome Project places a high importance on the autonomy of individuals and their right to take risks. Because of this, we require participants to be capable of making such a decision, and to demonstrate that they understand the risks related to participation by responding to exam questions. Individuals not capable of self-determination (due to youth or mental impairment) cannot participate in the Personal Genome Project; a caretaker or guardian cannot give consent on their behalf to participate in this study.

2. Beneficence

A study should maximize potential benefits and minimize potential harm for its participants. It is sometimes the case, however, that the subjects themselves are unlikely to benefit from a study -- for example, they may have an incurable condition the researchers seek to prevent. Thus, beneficence can also represent a benefit to a group of individuals that those subjects represent, rather than the subjects themselves. Sometimes this leads to ambiguous situations where subjects are exposed to risks with little chance of benefit, in the hopes of benefiting others -- the appropriateness of such studies depends on the risks and benefits involved, and on the autonomy of the subjects who decide to participate. Ethics commissions (notably Institutional Review Boards) must weigh the risks and benefits when giving approval to studies. During the course of a study, Data Safety Monitoring Boards watch over the safety and well-being of the participants.

Although there are some instances where whole genome data has benefited individuals, the Personal Genome Project believes these will be (at least at first) exceptional cases. In particular, the public release of genome data involves numerous risks and is unlikely to provide any immediate benefit to participants. However, creating such a public data set is vital for researchers and clinicians to learn how to use human genome data and use it to improve human health. As such, participation in the Personal Genome Project is best seen as a benefit to humanity as a whole, rather than to the participant themselves.

3. Justice

Who should benefit from a study, and who should bear the burden of risk? Fairness in distribution of benefits and risks means that research studies should not be exploitative. Particular groups should not suffer an undue burden of risks simply because they are more readily available or vulnerable to pressure, and groups which take risks should also represent groups which are likely to benefit from the research.

Participants in the Personal Genome Project have so far been a self-selecting group who have learned about our project through word-of-mouth or articles which mention the project. While we plan to extend our recruitment to individuals with specific diseases, we do so in the hope that better understanding and treatments for those diseases may arise from such study. We also hope to extend our recruitment to older individuals, who are able to share a lifetime of health information -- many older people are highly motivated to contribute to a better understanding of health and aging.

What is informed consent?

Informed consent is a process whereby individuals assess their willingness to voluntarily participate in a research project, based on their understanding of the purpose of the project. Presenting individuals with information regarding the risks and benefits of a study is necessary for their decision to participate, and it is important that this information is conveyed in a manner that fosters comprehension.

These study sections and exam questions are intended to ensure that you are presented with the information you need to make an informed decision regarding participation in the Personal Genome Project. However, the informed consent process shouldn't be just a one-time event. It should be an ongoing discourse that lets an individual assess whether to participate, before the research begins, and whether to continue to participate as the research progresses.

What is open consent?

“Open consent” is a type of informed consent developed in response to the highly personal and re-identifiable nature of genome research. Genome data is both sensitive information and difficult to make anonymous: it uniquely identifies an individual, and it has the potential to predict a variety of personal traits and medical conditions.

Traditionally human subjects research has made assurances regarding the confidentiality of data, as data was “anonymized” to protect privacy. In genomic studies participant data can be protected, but privacy and anonymity cannot be guaranteed. Even if precautions are taken, data security can be breached in many ways and assurances of anonymity are an over-promise.

Thus, participants in the Personal Genome Project are explicitly *not* promised anonymity. Our project recognizes that the public release of personal data and samples will expose participants to the risk of re-identification, and to other open-ended risks associated with the unrestricted sharing of data and samples.

Not all risks are known. This enrollment exam and the full consent form discuss some of the risks involved with participation -- including many hypothetical scenarios -- but there may be other risks we have not anticipated.

Module 2: Genetic Concepts

Introduction

We don't require expert knowledge in genetics to join the Personal Genome Project, but knowing a bit about the subject is important because it provides a context for understanding the type of data you will be receiving and making public. Genomes and DNA in particular are often regarded by many as powerful and mysterious -- after all, they are core parts of what make us who we are.

We hope to demystify the subject a bit by covering some topics we think are important for participants. And a quick note to the genetics experts in the audience: We know this brief review makes many simplifications, biology is a science full of exceptions. Sometimes we sacrifice technical accuracy for the sake of clarity and understanding.

DNA is the molecule of information

Deoxyribonucleic acid (DNA) is a molecule whose main purpose is carrying and copying information. The four "letters" of DNA -- A, C, G, and T -- represent the four different molecules forming links in the chain (these molecules are called nucleotides, and are named: adenosine, cytosine, guanine, and thymine). DNA is double stranded: two chains twisted around each other, zipped up as the nucleotides link together (A links with T, and C links with G). DNA can be copied by separating those two strands and creating a new partner.

DNA's instructions are used to create proteins, and these proteins are what carry out the various tasks need for cells and bodies to live. The translation of DNA to create proteins is called the "genetic code", and a section of DNA coding for a protein is called a "gene". Changes to DNA might change the way a protein functions (sometimes completely breaking it), or it might change how strongly the gene is "expressed" (in other words, how many copies of that protein it produces).

How do DNA instructions create a person?

How does this all come together to build a person? You can imagine how it may be similar to instructions for making a house. It might read something like this:

- * Using part 385 (a saw), cut a 10 foot long 2x4 out of pine, call this “part 16552”.
- * Place the end of part 16552 perpendicular to the midpoint of part 3550 (a 20 foot 2x4).
- * Use part 366 (a screwdriver) to insert four copies of part 6080 (screws) through part 3550 and part 16552.

Each of these parts could be thought of as proteins: some of them form structural elements, others interact to attach and modify the parts. That’s not all -- proteins are also reading the instructions, finding and using other genes to make new proteins.

In total there are around 23,000 genes that create proteins. When a sperm from the father fertilizes an egg from the mother, each contributes one half of the genome of the conceived child. This fertilized egg divides repeatedly along with its genome giving rise to all of the cells of that individual. Thus, all of the cells in your body carry a copy of your genome. What makes each cell of your body different is that they follow different instructions: a neuron has one set of active genes, a skin cell has a different set. Unlike the instructions you might have written, your genome isn’t in order -- the first instruction might be on page 680, the second on page 23, and the third on page 1,550!

Just as there is no single instruction “build a bedroom”, there is no single instruction in the genome to “build a hand”. Instead each gene is specifying a part of the whole. Over thousands of generations and across a population of billions of people, imperfections in the copying process and other environmental factors have introduced small changes in our genomes that make all of us unique and different from one another - even slightly different from our parents. Sometimes a change to these instructions has a drastic effect -- perhaps none of the windows have glass, because the “glass” part is broken! Other times the effect is subtle and normal -- for example, if the windows in your bedrooms were all built to be 2 inches wider. We refer to these differences -- both drastic and subtle -- as “genetic variations” or “variants”. (We prefer to use the more neutral term “variant” instead of “mutation” because the effects of many variants are unclear.)

Not all changes to DNA affect the instructions. In fact, very few changes do, because most changes to DNA occur outside of protein-coding regions. Genes are scattered throughout the genome: the parts that specify protein account for only 1% of the DNA. (If your genome were a book, the publisher is terribly wasteful of paper!) As a result, most changes to DNA have no effect on your traits -- but they are inherited nevertheless and provide a unique pattern that can be used to understand ancestry.

Nature versus Nurture

Genes are not destiny -- while all these proteins are working hard to create a person, a lot of what you are is influenced by the environment. This dichotomy is often expressed as “nature versus nurture”: the DNA is the “nature” (your apparent destiny) while the environment surrounding you is the “nurture” (the things outside the DNA that influence how you grow and change). The truth is that it is all an interaction: your genes are working together with their environment to produce a person.

This means that not all genetic variations predict absolute consequences, and their effects often depend on your personal history. For example, there is a genetic variation that predicts an increased risk of emphysema for people who smoke -- if you never smoked, then the effect of this variation is never seen. The environment you experience throughout your life influences the consequences of genetic variations.

Your “environment” starts early in life -- before you are born! From the moment of fertilization, there are variations in what people experience during their fetal development. Other variations in people could simply be random: a cell just happened to move to the left or right, in these early stages of embryonic development. What does this mean? Not all immutable traits are necessarily genetic. Some congenital aspects of ourselves (things we were born with) could instead be more related to random chance, or the very early environments. Who we are is a mixture of genetics, environment, and chance.

One copy from each parent

One last technical detail that is important for you to know about is that, for most genes, you have inherited two copies -- one from your mother's genome, one from your father's genome. Generally this works quite well, because you have a "back up": one of the genes might be defective, but the other one works fine -- you'll never notice any problem.

As a result, for many traits and diseases someone is only affected if they receive two broken copies of a gene -- one from each parent. These traits and diseases are called "recessive" because they "recede" in the presence of a working version of the gene. The parents of the affected individual could be totally unaffected if they only have one broken copy each (they are "carriers" of the genetic variation), and their chances of a child being affected is 25%. This is because there is an equal chance that each parent will pass on either the good copy or the broken copy (it's like flipping a coin twice: both of them got "tails" and passed their broken version of the gene to the child).

There are other traits and diseases where the genetic variation causes an effect even when seen only once ... it might even create an additional function -- instead of a broken screwdriver imagine one that keeps trying to use nails as well as screws, destroying many nails in the process! A person only needs one copy of such a variation for it to have an effect. In this case the variant is called "dominant" because its qualities "dominate" regardless of the other copy. When a parent has a single copy of a variant, their child has a 50% chance of inheriting it.

Finally, many variations may have intermediate effects -- one copy has a little effect, two copies have a larger effect. Others only have an effect when the person *also* has a different, additional variation. When looking at some traits, all of these intermediate effects can overlap in a complicated way -- many common human traits (like height and skin color) are called "multigenic" because they are affected by genetic variations in many different genes.

In summary, just because someone carries a variation associated with a disease doesn't mean that they will have the disease, or even that they have an increased risk. It depends on how strong the variant's predicted effect is, how that effect interacts with the environment, and whether that effect is recessive, dominant, or something more complicated. It is also important to remember that our understanding of genetic variation is far from complete -- scientific literature can sometimes be mistaken regarding the true impact of a genetic variation.

Module 3: PGP Enrollment Procedures

Introduction

The Personal Genome Project (PGP) seeks to publicly share genome data combined with public health and trait information. This module describes the procedures involved regarding your enrollment in the project.

In addition to exposing yourself to various risks, you will also expose your relatives to many of these same risks because they share DNA with you. Participation is ultimately your decision, but we strongly recommend you discuss participation in the PGP with your relatives before you submit your enrollment application (and on an ongoing basis throughout your participation).

PGP enrollment steps

The steps of PGP enrollment are as follows:

(1) Mini-consent and account creation	<p>You have already performed this step. This small consent form allows us to ask you questions regarding your eligibility, and some other demographic and trait information. Personally identifying information provided here is considered private.</p> <p>You should have signed this form electronically with your name and email.</p> <p>You may not sign this consent for another individual nor perform later enrollment steps for another individual, even if you have legal authority to make health care decisions on their behalf (e.g. a guardian). The decision to enroll must be made by individuals capable of giving their own consent.</p>
(2) Email address verification	<p>Your email address was verified in this step. Your email is a critical aspect of your account: resetting your password will send a new password to your email account. Anyone who has access to this email address will be able to access to your personal PGP data and could pretend to be you.</p>
(3) Eligibility screening	<p>A small set of questions to see if you meet our eligibility requirements.</p>
(4) Enrollment exam	<p>This enrollment exam was created to ensure that you are informed of the risks and benefits of participation in the PGP.</p>
(5) Full consent form	<p>After passing the entrance exam, you will be presented with the full consent form. We encourage you to read it carefully -- signing this form means that you agree to the protocols and risks associated with the PGP.</p>
(6) Designate proxies and enter initial data	<p>As your account starts, we ask for some initial data. One particularly important piece of information we ask you for is who you designate as proxies: people who should act in your stead to make decisions regarding your participation in the PGP, should you die or become mentally impaired. The identities of these proxies will be considered private information.</p>
(7) Submit enrollment application	<p>Generally all individuals who pass these steps and submit an application become enrolled in the PGP.</p>
(8) Active enrollment	<p>Enrollment should occur within 1-2 months of submitting your application; you should receive an email notification. Once enrolled, you will be able to log in as an active participant and perform associated activities on the website (including addition of data to your public profile and signing up for sample collections).</p>

Remember that your participation is always voluntary. You may withdraw at any time, for any reason.

In addition, the PGP may decide to end your participation at any time, whether or not you agree to such a termination. Some reasons this may occur include: if we have reason to believe a participant misrepresented their identity, if we believe a participant was enrolled by someone else (i.e. not autonomously), or if we suspect a participant is likely to submit samples or data belonging to other individuals.

Module 4: PGP Participation, Specimen Collection & Data Release Procedures

Introduction

This module describes what you can expect to occur once you are enrolled, including: how we will collect samples and data from you, how these things will be become public, and what long-term commitments you are making as a PGP participant.

Participation is an ongoing relationship

Participation in the PGP is an ongoing relationship: our understanding of genomes is new, and so we are likely to revisit participants for follow-up studies, questions, and specimen recollection. By participating in the PGP you are consenting to being recontacted by us. In many cases responding to our follow-up requests is voluntary. There are two types of ongoing follow-up that is required of all participants: Quarterly Safety Reports and updates to the consent form.

Quarterly Safety Reports are our way of monitoring participants for adverse outcomes. Four times a year you will be emailed a reminder asking you to answer some brief questions regarding your current status, and whether you have any developments to report related to your participation in the PGP. Your responses to these reports are confidential and will be kept as private data.

You may also be asked to sign updates to our consent form. Because participation is ongoing, our consent document is also an ongoing process -- as the project evolves we may need to periodically revise the consent form accordingly. After a change to the consent form occurs, all participants must approve of the new consent form to be considered active participants.

Failure to complete quarterly safety reports or to approve the new project consent result in account deactivation (but completing the overdue report and/or reviewing and accepting the new consent will reactivate your account).

Ongoing participation also means you may be recontacted for various reasons, including: updates to our website, new specimen collection opportunities, and other PGP-related announcements.

Specimen collection

We are pursuing a variety of methods for specimen collection, including saliva and blood collection. We may ask you to provide your contact information (e.g. address and/or phone number), which will be kept private. Because of the resources necessary for specimen collection and processing, we may not be able to collect a specimen from you right away. Specimen collection methods should pose minimal risk to you, but will not necessarily succeed - you may be asked to contribute a second specimen, or we might later conclude that a given collection method has yielded unusable specimens.

Specimen collections are always voluntary; you are not required to participate in any specimen collection events. As part of providing us with a specimen, you are consenting to an open-ended set of things which may be derived from it. Data generated from those specimens will be shared both with you and publicly. Such data could include: genome sequencing data, epigenetic data, protein and chemistry assays, and profiles of bacteria and viruses in your samples.

Release of genome data and other sample-derived data

Currently, when we derive new types of data from specimens (e.g. whole genome sequencing data) we give participants a 30 day period of private access to that data. For whole genomes, in addition to raw data access we will give you our current interpretation of your data -- this interpretation is part of our research development and is extremely incomplete, it is almost certainly missing some variants and misinterpreting others.

We will notify you of the new data, but we will not require any approval from you to release the data -- the data will be added to your public profile 30 days after we have notified you. You may choose to withdraw from the PGP during this period and have all existing public data, specimens, and specimen-derived samples removed, in which case this private data will not be added to our public database.

Only new types of data will be held private for a 30 day period. For some types of data it is likely we will improve upon the data as our methods for laboratory and computational analysis improve. Updates that improve upon the quality of a given type of data may be made public immediately. For example, if we reprocess your whole genome sequencing data to correct or add information, those improvements may be made public immediately. Updates to data interpretations (e.g. genome interpretations) may also be made public immediately.

Remember that analyzing specimens to generate data will often take some time: whole genome sequencing currently takes months, and not all specimens can be processed immediately due to our limited resources. When possible, we will try to update you regarding the status of your specimens. We do not guarantee that the specimens you provide us will be analyzed, and even if analyzed they may fail to produce data.

Creation of cell lines

For some types of specimens (include blood samples) we may derive cell lines. Cell lines are a renewing resource that is produced by growing your cells within the laboratory. Such a resource allows others to more easily access copies of your genomic material. Other groups could potentially use the cell lines to replicate or improve our sequencing results or to perform various scientific studies regarding the function of variations within your genome. Although there may be some regulations or agreements involved to acquiring cell lines, we will try to share these as freely as possible by making them available in an established repository.

Health records and other user-provided data may be immediately public

Data can also be added to your profile by you, directly. This may include: addition of information regarding your health and traits, uploading of genetic data files acquired elsewhere, and making a public association of your name with your profile. All of these additions may be immediately added to your public PGP profile.

Health record data and other trait information is sensitive information, but it is also extremely valuable for science. Genome researchers seek to understand the connection between genetic variations and various health conditions and traits. Participants in the Personal Genome Project who share health information are especially valuable because they enable researchers to compare methods and results. We will provide you with opportunities to respond to surveys, import electronic health records, and upload various files.

Providing us with these data is completely voluntary. Such data can be sensitive and may greatly increase your risk of being re-identified. You should review such data carefully to ensure you are not inadvertently disclosing personal, health or insurance information you do not wish to be made public. What you make public is at your discretion, you are not required to provide us with any of these items to be a participant in the PGP.

Module 5: Risks & Benefits Part 1:

Potential consequences of receiving personal genome data

Introduction

Much of the data produced by the Personal Genome Project -- including whole genome data -- is produced using newly developed research techniques. Such techniques may produce erroneous results, and the interpretation of such data is in many cases ambiguous and rapidly developing. Whole genome data also has the potential to reveal much about you as a person, and this may include uncovering unexpected and potentially upsetting information.

We outline these risks involved here, but there may be other risks we have not yet anticipated. In particular, we plan to collect and analyze samples to produce data other than whole genome sequences (e.g. microbiome, epigenetic, or protein data). Most of the risks we discuss in this and the next module regarding your genetic data (e.g. errors in interpretation, potentially upsetting findings, potential discrimination) also apply to these other types of data.

Errors in data

No test is 100% accurate, and this is especially true of data generated by research studies. Whole genome sequencing is, on the whole, an amazingly accurate technology. We estimate the chances that any given genetic variant has been called in error is less than 0.5% (1 in 200).

While that error rate is quite low, it's still high enough to be troubling. This is especially true for when we find variants which are themselves quite rare in the population. If a DNA variant is only seen with 0.5% frequency, then the chances of you having the variant and the chances of an error are roughly equal. Variants with serious clinical effects also tend to be the rarest, and so often the scariest findings are also the least trustworthy.

Errors may occur in other ways too -- from sample mix-ups to data processing mishaps.

We make efforts to reduce such errors, but our research data is far from "clinical quality" and should never be considered a substitute for clinical testing. Before acting upon findings, you should always have them confirmed by a licensed healthcare professional.

Unanticipated ancestry discoveries (e.g. non-paternity)

Even a small amount of genetic data is enough to make many inferences regarding a person's ancestry. These discoveries may occur as the PGP develops tools for ancestry analysis in coming years, or they may be discovered by yourself or third parties that study the public data.

If this data is compared to data from relatives, you may discover that you are not related to people you thought you were related to. A common example of this is non-paternity: discovering that an individual's father was not their biological father -- something that may be a surprise to the father, the child, and sometimes even the mother. Barring accidental sample mix-ups, such a discovery can be inevitable and is not sensitive to sequencing errors. Other examples could be discovering a similar non-paternity finding in a parent & grandparent (uncovered when your data is compared to a cousin), or discovering unknown cases of non-maternity or adoption.

More distant ancestry discoveries may also be discovered as your genetic data is compared to population data. For example, you may discover you have racial heritages that differ from what you had thought.

Unanticipated findings predicting serious clinical effects

One negative consequence of genetic testing may be the possibility of finding out information about serious medical conditions. It may be something you can do little or nothing to prevent. It may be something rare and ambiguous, where your prognosis is very uncertain. These discoveries can lead to anxiety and stress, and they can be very difficult to anticipate.

Sometimes you may have reason to be concerned about a particular finding. For example, an individual whose parent has Huntington's Disease will be concerned about discovering that they carry the same mutation, or an individual whose parents both have Alzheimers Disease might be especially concerned about their own genetic predisposition to the disease.

Other discoveries, however, may be totally unanticipated. This case is best illustrated by the experience of a recent PGP participant found to carry JAK2-V617F in his blood cells -- a mutation associated with causing rare and serious blood diseases. This is a mutation that probably occurred and accumulated in his blood cells during adulthood -- it was not inherited, and therefore something family history could not have predicted.

Ambiguous findings

Not all genetic variants are well understood. In fact, many or most of them are quite poorly understood. Does this genetic variant really cause a disease? Is there a 100% chance? How severe will it be?

This is especially true of the rarest variants -- variants which may have one or two publications implying they cause a disease, but which are so rare that they have very little evidence. With so few observations it can be hard to know whether the variant really is associated with a disease and what the prognosis is.

Failure to report serious findings

Not all serious findings will be uncovered immediately. Some genetic variants are not successfully tested using our current genome sequencing technology. Huntington's disease is one example of this: because of how the variant is formed in the DNA, it is very difficult for current whole genome sequencing techniques to distinguish between disease-causing and benign variants in the gene. As we improve sequencing methods we expect this variation to be discovered and reported -- but it is important to be aware that current technology has limitations.

In addition, many findings will be uncovered later as our ability to interpret improves. Currently our interpretations of genome data are extremely incomplete. Even as we work hard to fill in the gaps in our interpretations, the knowledge of genetic variants is constantly growing. It is very likely that serious findings will be made in your genome data long after you first received it, and after it has been made publicly available.

Flaws in interpretation

Finally, even if a variant is reported to have an association, there may be a flaw in our interpretation of existing literature. Such human error is inevitable when reviewing hundreds of publications and, while we hope the errors are corrected over time, new literature is always being published. In addition, the literature itself may contain errors. Our interpretations are extremely new and essentially untested. We provide them because they give some insight into how the world may view your genome data, but whole genome interpretation is a very new field of research and extremely prone to error.

Potential financial costs

The results produced by the Personal Genome Project cannot be used as a medical test. Nevertheless, findings in your genome may suggest you should have clinical follow-up -- e.g. clinically-certified diagnostic testing or medical advice. As a research project, the Personal Genome Project is not able to provide you with these services and will not cover the costs. In addition, it is possible your health insurance company will not cover these costs.

As a result, you are potentially risking the financial burden of follow-up diagnostic testing -- even if our finding turns out to be an error in sequencing or interpretation.

Module 6: Risks and Benefits Part 2:

Risks associated with making your data public

Introduction

One of the core principles of the Personal Genome Project (PGP) is that scientific material should be shared as freely as possible. This means that the PGP will make your genome data public (and any other data derived from the specimens you provide us). In addition, if participants choose to publicly share information regarding their health and traits, this information will also become publicly connected to their genome data. Finally, we will also try to share anything we derive from specimens as freely as possible (e.g. lymphocyte cell lines derived from blood samples).

This sharing is a great benefit to the scientific community, but it has many potential risks to our participants. We try to outline the major risks we think you are taking, but not all risks are mentioned here -- indeed, some risks may arise from future technology developments that are currently impossible to anticipate.

Loss of privacy

The most obvious risk you have is that your data will become associated with your identity -- someone might, for example, figure out that genome #982 belongs to you. This risk seems so likely that many participants decide to associate their names with their genomic data at the outset (although this is not required).

Imagine, for example, that you wish to tell your friends about a particularly rare variant you have: only 1 in 200 people have this interesting genetic trait. From this information alone your friends might be able to figure out which genome is yours: your sex will narrow it down by a lot, your race narrows it down more, your hair color or last name might be additional clues... before you know it, they've figured out exactly which genome is yours! Suddenly they are able to learn many personal things about you, including everything predicted by your genome and any health record data you've publicly shared in your PGP profile.

Even if you are perfectly good at keeping secrets, your privacy could be lost as third parties become especially skilled at analyzing our data and inferring identity from it. Identities can be discovered with surprisingly little information -- for example, the combination of sex, birth date and ZIP code is specific enough to be uniquely identifying information for 87% of people! This means that some could associate your identity with your PGP data, without your knowledge and without your consent.

Discrimination on the basis of genetic or health data

Politicians have been passing federal and state laws that try to protect individuals against genetic discrimination, but these protections are untested and incomplete. In 2008 Congress passed the Genetic Information Nondiscrimination Act (GINA), designed to protect individuals from the dangers of genetic discrimination. GINA prohibits health insurance companies from denying coverage or charging higher premiums solely on the basis of genetic information. GINA also prohibits employers from using genetic information for hiring, firing, promotion, and other employment decisions.

However, there are significant types of discrimination that GINA does not protect against. For example, GINA does not prohibit discrimination for other types of insurance (e.g. life insurance, long term care insurance, and disability insurance). In addition, as with other types of discrimination, it could be extremely difficult to prove discrimination has occurred. You might never know whether your employer found your PGP data and read about your genetic findings. GINA and similar laws are new, incomplete, and relatively untested -- you should not assume that your genetic information, if it should become associated with you, would never be used against you in a way you found objectionable, whether or not prohibited by law.

Finally, GINA and its state counterparts generally do not restrict the ways in which others (including employers and insurers) may use certain non-genetic data about you that may become publicly available through the PGP. Information that you share with us about your personal health and traits are extremely valuable for the scientific understanding of genomes. Unfortunately, this can also represent sensitive information (e.g., a personal or family history of mental illness). While GINA and other laws, including the Americans with Disabilities Act, prevent certain uses of certain information, you should carefully consider the information you make available and how it might be used against you or your family.

Once public, it may be impossible to completely remove your data

Once we make your genetic data public, other people can download the files. In addition to analyzing that data they could choose to share those files with others, host them in their own database, or publish derivative data. You may withdraw from the PGP at any time and choose to have your public data removed from our database. However, others may already have copied this data and we cannot control what they do with it.

Discovery of serious findings after data is already public

As described in Module 5, our ability to interpret genomes is incomplete. It is likely that important interpretations of your genome data will be made after your data has already been made public. Potentially sensitive information may later be interpreted after the genome data has been public, and it may be too late to remove it from non-PGP databases.

Unanticipated uses of your data and cell lines

The list of potential uses of your data and cell lines by other individuals is diverse and sometimes worrisome. The benefit of these things is that other researchers will use them in their own work, greatly facilitating the process of scientific research. Other researchers might also create their own interpretations of your genetic data -- and these could make incorrect claims regarding your predisposition to traits and diseases that we cannot control. Someone might match your public data against other genetic databases to find matches for yourself or relatives - this includes criminal and forensic DNA fingerprinting databases as well as other genetic research studies.

More nefarious uses are also possible, if unlikely. DNA is commonly used to identify individuals in criminal investigations. Someone could plant samples of DNA, created from genome data or cell lines, to falsely implicate you in a crime. It's currently science fiction -- but it's possible that someone could use your DNA or cells for in vitro fertilization to create children without your knowledge or permission, or to create human clones.

Not all risks are known -- there may be some risks that are impossible to anticipate at this time. It is very difficult to predict what future technologies and techniques could be applied to your DNA sequence data and cell lines.