Lab: The Human Genome Project (HGP) and Personalized Medicine

Anya Goodman, Ph.D. Department of Chemistry and Biochemistry, Cal Poly – SLO
James Youngblom, Ph.D. Department of Dept. of Biological Sciences, CSU-Stanislaus

Introduction

"Do I have higher than average risk for developing lung cancer? …breast cancer? …colon cancer? …alcoholism? …side effects from a particular medication?" The answers to these questions are becoming accessible and affordable to an average citizen. The explosive growth of biological data brings the promise of personalized medicine: healthcare customized to each person's genome. Lab participants will use web-based tools and databases to explore gene variation that changes patient's response to medication and risk for developing diseases. Both scientific resources (NCBI) and commercial site targeted to lay people (23andme) will be used in the exercise that assumes minimal prior knowledge of genetics.

Goals: the goals of this exercise are to introduce participants to the analysis of DNA sequences using computers (bioinformatics) and to assess the effects of the Human Genome Project and bioinformatics on the health information available to scientists, doctors and general public.

Background

Partial instructions for making a living organism are stored in the organism's genome (a set of DNA molecules carrying information). DNA's alphabet contains only 4 letters: A, C, T, and G. Human genome contains about 3 billion letters stored in 23 chromosomes. If human genome was analogous to the collected works of a famous author, here is how they would compare:

	Collected works of Fyodor Dostoyevsky	Human Genome
Original language	Russian: 33 letters	DNA: 4 letters A, T, C, G
Translated into	English: 26 letters	Protein: 20 letters
"Printed" in	30 volumes	23 chromosomes
"Meaningful" content	32 novels and short stories	~25,000 protein sequences (genes)
Total # of letters	?	~3 billion
% meaningful letters	?	1%

Sequencing the human genome means finding out where all of the 3 billion nucleotides of the human genome are located. It turns out that the longest human chromosome, chromosome #1 contains ~247,000,000 nucleotides, whereas the shortest one (chromosome #21) contains ~47,000,000 nucleotides. We know the order and position of the nucleotides of a typical human genome (human genome draft completed in 2001, nearly complete genome published in 2003), but we still have to figure out what they all mean and how living things work. To do this, we need to be able to compare DNA sequences. Sequence comparison is greatly facilitated by using computers. Before we practice on the computer, we'll run through a paper/pencil exercise.

Instructions: please, follow the directions and hints in this worksheet and answer the questions on the separate answer sheet.

Part 1. DNA sequence comparison. You are given a fragment of DNA sequence from a patient, Greg Mendel, who wants to know whether he should be treated with beta-blockers (like metoprolol) to prevent heart attacks or to treat a heart attack. We will first compare patient's sequence (query) to a sequence database to see if there are matches; then, we will look carefully and determine whether the match is exact or if there are any differences between sequences.

Query: Greg Mendel's sequence
Database: fragments of sequences from known genes

Find what sequence in your small paper database **best** matches the patient's sequence.

1A. What is the best match to the query in the paper database? # _____.

Compare the best match from the database to the patient's sequence and circle all letters that are different between the two sequences. Do not worry if the beginning or the end of your sequence are missing some of the letters; we are looking at partial sequence and do not know whether unspecified sequences match.

When a particular gene has different "spellings" in different individuals, the different versions of the gene are called **alleles** of that particular gene. A difference that is just a substitution of a single letter for another one is called a **SNP (single nucleotide polymorphism, pronounced 'snip').** Many studies trying to find the genetic basis of disease try to correlate disease state with a particular **SNP** or **allele.**

1B. Does the patient's DNA sequence contain any SNPs in the given region? Yes or **No**
If yes, how many? _____

Our paper database does not have enough information to answer the patient's question about heart attack treatment, so we will now turn to the scientific database to repeat our search.

Part 2. Our next goal is to figure out what gene does the patient's DNA sequence came from. We will search public DNA sequence database GeneBank used by scientists all over the world and housed at the National Center for Bioinformatics Information (NCBI). The search will require the use of computer, because the database is really large. As of June 2013, GeneBank contained 152,599,230,112 bases in 165,740,164 sequences. To see how many sequences and bases it currently has you can look for GeneBank statistics at

http://www.ncbi.nlm.nih.gov/genbank/statistics

We will use a computer program, BLAST (Basic Local Alignment Search Tool) to search the database for matches to our query sequence. The result will be similar to what you found in Part 1: a matching record from the database and an alignment between the database sequence and the patient's sequence. The main difference: the database record we find will be linked to other

databases containing information we need (e.g. gene name, gene function, SNPs and alleles, scientific literature about the gene and connections to prevention of heart attacks etc.).

Instructions for BLAST:

A. Open the home page for the National Center for Bioinformatics Information (NCBI- www.ncbi.nlm.nih.gov)

B. Find BLAST under "Popular Resources" on the right-hand side.

C. From the main BLAST screen, choose "nucleotide blast" (see figure 1 below).

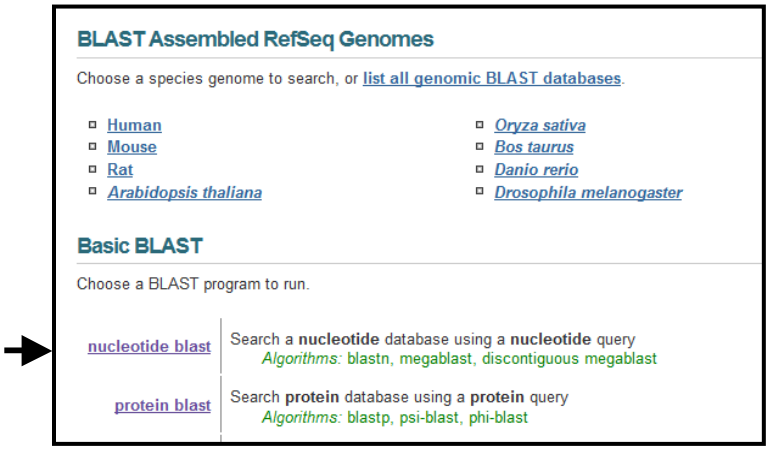


Figure 1. Screen shot of BLAST main page.

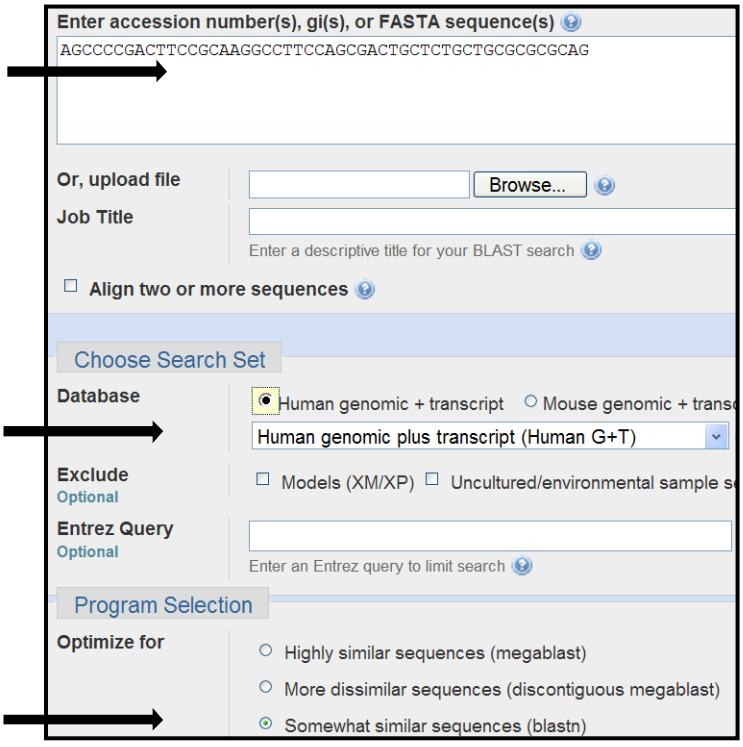


Figure 2. Screen shot of BLAST entry page. Three arrows show settings for steps D, E, and F.

D. Type in (or copy and paste) your query sequence from Part 1 into the window titled "Enter accession number(s)..."
E. The database we will be searching is "human genomic + transcript" (see arrow in fig 2).
F. Scroll down to "program selection" area and change program to "somewhat similar sequences (blastn)."
G. Click on the "BLAST" button at the bottom and wait for the result.
H. Interpret the BLAST output. There are three areas we will examine to answer the questions below: graphic output, table output and the actual alignments.

Graphic output: our query is represented as horizontal line from nucleotide 1 to 53 (shorter if you did not have the patience to enter the whole sequence). The **matches or hits** from the database are shown as horizontal lines below. Color represents how similar the matches are, length shows the part of the query that matched. You can see the names of the hits by rolling the pointer over the lines.

Table output: list of hits along with quantitative parameters of how good the match is. There are two numbers we really care about: **Max score** and **E-value.** The higher the score – the better our match is. Default BLAST algorithm assigns +2 for each matching nucleotide and -3 for each mismatch to get a raw score, and then adjusts that score by length and database size to calculate the "**Bit score**." **E-value** (or expect-value) estimates how many times do we expect to find a match of the same quality (bit score) in the database purely by chance. The lower the E-value – the better. The matches with E-value above 1 can be discarded.

The alignment (figure 3)**:** the first line with ">" gives the address (accession number) of the full entry in the database, name of the entry, and length. Summary section below the name shows bit score and raw score in parenthesis, E-value and % identity. The last part is the alignment itself: query on top, subject (sequence from the database) on the bottom, and vertical lines for each nucleotide match between them. The numbers specify nucleotide position in each sequence.

Link to data base record

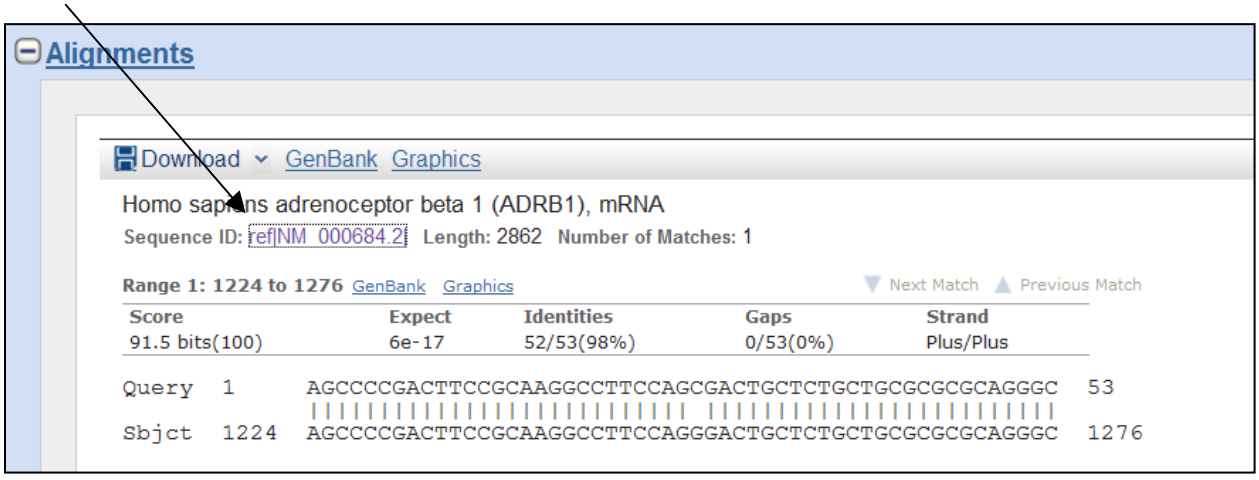


Figure 3. Alignment of patient's sequence to the database record ref|NM_000684.2|

Examine your alignments and answer the following questions:

2A. How many significant hits did we find in the human database? (E-value below 1) _____
2B. What is the name of the gene? _____What protein does this DNA code for? _____
2C. What chromosome does this gene reside on? _____
2D. Why is there more than one meaningful match to our sequence in the human genome?

Hint for 2D: the first hit on the list is "mRNA," not genomic DNA. RNA is a molecule similar to DNA and we can think of it as a working copy of the gene, before the gene gets translated from the DNA language into the protein language. If you ignore the mRNA, there are still multiple genomic DNA matches(hits). Do they all match our query equally well?

Part 3. Investigating patient's Single Nucleotide Polymorphism (SNP) and patient's genotype.

In part 2, you have identified the patient's gene that contains a SNP. This SNP is in the coding region of the gene and affects amino acid #389. To find out whether the difference in one nucleotide changes the amino acid sequence of the protein, we will need to translate DNA sequence into the protein sequence.

Use the genetic code table (look up on the web) to translate your DNA sequence into the protein sequence. Start translation at the first nucleotide. Note, the first amino acid is not methionine because we are looking at a fragment of DNA sequence, not the whole gene.

```
Sequence 1…AGCCCCGACTTCCGCAAGGCCTTCCAGCGACTGCTCTGCTGCGCGCGCAGGGC
Amino acid:
or
Sequence 2 AGCCCCGACTTCCGCAAGGCCTTCCAGGGACTGCTCTGCTGCGCGCGCAGGGC
Amino acid:
```

3A. Compare your results with your lab partner's:
What codon (3-letter code) is different between two sequences?
Sequence 1 _____ Sequence 2? _____
This codon codes for amino acid #389 in the adrenalin receptor.
What is the amino acid #389 in sequence 1? _____In sequence 2? _____

3B. Do you expect this specific change in amino acid to affect the structure of the protein? Briefly explain why.

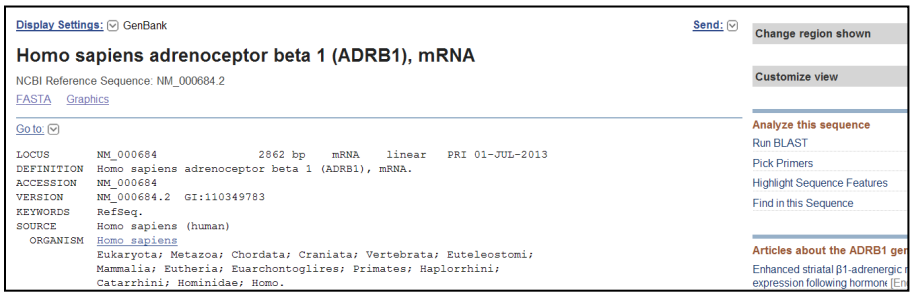
3C. Both sequence 1 and 2 came from the same patient (Greg Mendel) and match the same gene. Why does the patient have two copies of this gene? _____
What is Greg's genotype at this SNP? Circle one: CC CG GG

Next, we would like to understand the connection between patient's genotype and phenotype (efficacy of heart attack treatment with beta-blockers). This information is stored in OMIM

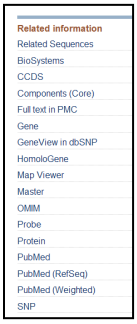
database at NCBI. OMIM database contains records devoted inherited diseases and contains summaries of scientific and medical literature. Navigating through the information requires advanced biology training.

Optional: If you want to check out the scientific database, follow these steps:

- From your BLAST output to gene/transcript record (ref|NM_000684.2; figure 3) click on the link to the record – you will see a page with a lot of information that looks like this:



- Scroll down looking for the header "Related information" on the right hand side of the page. Click on link to "OMIM"



- Search OMIM for the gene symbol or gene name.

We are going to take a different route and look at a site geared towards general public/consumers.

Part 4. Genetic testing and health information for informed (for all?) citizens.

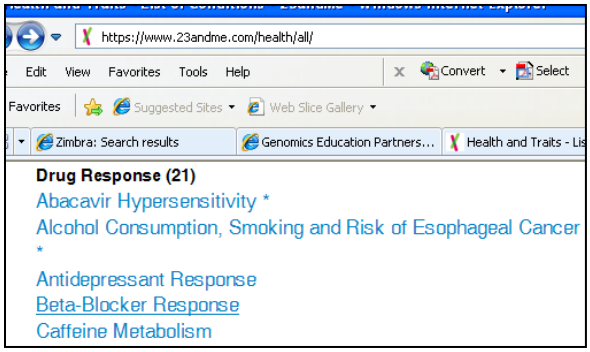
Prediction: in the near future, sequencing an individual's genome will cost less than $100 and become affordable and widely used in medicine. Currently, genotyping for ~1000 markers (SNPs) costs ~$100 and the body of knowledge about links between SNPs **(genotype)** and individual's characteristics **(phenotype)** is rapidly growing thanks to the awesome power of computers.

There are several large companies offering similar services. We are using **23andme** because this company has good education resources and recently updated their sample reports (instructor has no financial or personal ties to the company☺).

Open the 23andMe site with the list of all health related sample reports,
https://www.23andme.com/health/all/

scroll down to “Drug response,”

and choose “Beta-Blocker Response.”



Click on the link to the “Beta-Blocker Response” report, look through the example report and answer the following questions:

4A. Action of what hormone do the beta-blocker drugs (bucindolol, metoprolol, etc.) block?

4B. How do these drugs work in preventing/treating heart attacks?

4C. Propose a hypothesis to explain on the molecular level why patients with a “G” in a particular SNP do not respond to beta-blocker drugs.

4D. Should Greg Mendel be treated with beta-blockers? Yes or No Explain why:

This example illustrates one application of human genomics to medicine. The list of known SNPs associated with health-related issues is large and constantly growing. Next, we’ll try to predict our patient’s phenotype for two other two conditions. Go back to the list of all example reports (https://www.23andme.com/health/all/). Scroll down the list. Various characteristics (phenotypes) are grouped as following: carrier status, drug response, traits, disease risk. When you click on a sample record, you can get additional information about the role of genetics in a particular characteristic and known genotypes (SNPs/markers) linked to a particular phenotype. Note, there may be more than SNP for each trait.

4E. Diabetes predisposition: under “disease risk,” find “Type 1 diabetes” and “Type 2…”

Does Greg Mendel have lower than average risk of developing type 1 diabetes? Yes No

Does Greg Mendel have lower than average risk of developing type 2 diabetes? Yes No

How many SNPs were used to predict predisposition for type 2 diabetes? _____

Do all SNPs make similar predictions: all increased or all decreased risk of disease?

It is possible that there are other SNPs in the human genome that have not been discovered yet. Imagine, that Greg Mendel had his entire genome sequenced and knew **ALL** the SNPs affecting his predisposition to type 1 and type 2 diabetes.

For what type of diabetes would scientists/doctors be able to predict the patient's risk with more confidence? Type 1 Type 2 Explain.

If genetic information suggested lower than average risk of developing diabetes, could the patient still develop the disease? Yes No Explain

4F. What else can we learn about Greg Mendel from his genotype? Pick one trait that is of interest to you and summarize your findings on the answer sheet.

Part 5. Inferring phenotype based on genotype.

Table 1 lists genotype of three real people at several SNPs that have been linked to particular traits. Use 23andMe website to fill in the missing information in the table and predict the phenotype of these three individuals. You may work in small groups and divide up the traits to look up; then discuss your findings in the group.

Table 1. Genotypes and Phenotypes of three individuals:

Trait	SNP accession number	Summary: gene, coding/non-coding region, genotype-phenotype, etc.	X Genotype Phenotype	Y Genotype Phenotype	Z Genotype Phenotype
Lactose intolerance	rs4988235	SNP in the regulatory region of the gene coding for lactase enzyme. GG – intolerant, enzyme production is reduced after childhood, AA and AG- enzyme produced into adulthood, lactose tolerant.	AA Tolerant	GA Tolerant	GG Lactose intolerant
Height	rs6060371		GT	GG	GG
Height	rs1042725		CT	CT	TT
Eye color	rs12913832		GG	AG	AA
Eye color	rs12896399		TG	GT	TT
Eye color	rs1393350		GA	GG	GG
Hair Curl	rs17646946		GA	GG	GG
Hair Curl	rs7349332		CC	CC	TT
Longevity	rs2764264		TT	CT	TT
Longevity	rs2542052		CC	AA	AA

Hint: you can paste SNP accession number from the table (rs….) into the search box or browse the list of traits.

Table 2. Summary of predicted phenotypes and identification of three people.

Trait (phenotype)	X	Y	Z
Lactose intolerance	**No**	**No**	**YES, lactose intolerant**
Height (average, taller or shorter than average)			
Eye color (Blue, Brown, Green)			
Hair curl			
Longevity			
Name of the person			

The instructor will tell you the names of the individuals in the table – see if you can match X, Y, and Z to the names.

Part 6. Conclusions and wrap up:

Why is it difficult to predict traits from just knowing DNA letters at each SNP? Can you think of three reasons?

Despite the reasons above, you may be considering getting yourself genotyped. Think carefully about the impact on you, your siblings, other family members and all future offspring related to you!

There are several well-known personal genotyping companies: Navigenics, deCODEme, 23andMe, Pathway Genomics. All offer various types of SNP screening. Everyday the information that they provide becomes more accurate and informative. If you have a particular nagging personal genetic question, you might want to be tested now. Maybe you are concerned about a family history of breast cancer. Do your homework. Find out how many breast cancer mutations are screened before selecting a genotyping company. If your concern is prostate cancer, find out how many SNPs are in the panel for each company's prostate cancer test.

Individuals who have been genotyped, might want to visit the community-based wiki SNPedia. SNPedia allows anyone to post information on SNPs. Currently there is information on ~24,000

SNPs. Should you have your genome sequenced, you will probably want to do two things: 1) keep things in perspective, and 2) discuss your results with a qualified genetic counselor.

Personal genomics is about to undergo a huge transformation. As the cost of DNA sequencing continues to plummet, sequencing the entire human genome will soon become affordable, and individuals will be able to analyze the whole genome rather than selected SNPs. After many individual genomes have been sequenced, scientists can analyze each of the 3 billion nucleotides. This could be much more informative than analyzing a panel of SNPs. When the cost hits $100 or so for whole human genome sequence, be prepared for an onslaught of publicity.

Discussion questions:

1. According to A. D. Baxevanis of the Nat. Human Genome Res. Inst.: "the advent of the genomic era will have a profound effect on how health care is delivered from this point forward." List anticipated changes in healthcare in the post-genomic era. Discuss potential benefits and risks of applying human genomics to medicine.

2. According to neurologist Robert Green: APOE is the only gene for a common disease that "meaningfully increases someone's risk in a way that could conceivably mean something to an individual." (note- one copy of APOE4 means a 2-3 fold increased risk of Alzheimer's disease; two copies of APOE4 means about a 15 fold increased risk.) What point did Robert Green try to make with that statement? Do you agree or disagree? Why?

3. According to Muin Khoury of the Center for Disease Control and Prevention: Testing for twenty SNPs for diabetes and heart disease "doesn't tell you more or less than what you already know based on your family history and body mass index." Explain the point made by this statement. Do you agree or disagree? Why?

4. "We used to think our fate was in the stars, now we know, in large measure, it is in our genes." James Watson, *Time*, March 20, 1989. Explain the point made by this statement. Do you agree or disagree? Why?

5. If you were offered one free genotyping by 23andMe, which genetic test would you choose? Why?

The HGP and Personal Health Answer Sheet **Name_____**

The patient's sequence you are analyzing is SEQ1 or SEQ2 (circle one)

1A. What is the best match to the query in the paper database? # _____.
1B. Does the patient's DNA sequence contain any SNPs in the given region? Yes or **No**
If yes, how many? _____
2A. How many significant "hits" did we find in the human database? _____

2B. What is the name of the gene? _____
What protein does this DNA code for? _____

2C. What chromosome does this gene reside on? _____

2D. Why is there more than one meaningful match to our sequence in the human genome?

3A.What codon (3-letter code) is different between two sequences?
Sequence 1 _____ Sequence 2? _____
This codon codes for amino acid #389 in the adrenalin receptor.
What is the amino acid #389 in sequence 1? _____In sequence 2? _____

3B. Do you expect this specific change in amino acid to affect the structure of the protein? Briefly explain why.

3C. Both sequence 1 and 2 came from the same patient (Greg Mendel).

Why does the patient have two copies of this gene? _____
What is Greg's genotype at this SNP? Circle one: CC CG GG

4A. Action of what hormone do the beta-blocker drugs (bucindolol, metoprolol, etc.) block?

4B. How does the drug work in preventing/treating heart attacks?

4C. Propose a hypothesis to explain on the molecular level why patients with a "G" in a particular SNP do not respond to beta-blocker drugs.

4D. Should Greg Mendel be treated with beta-blockers? Yes or No

Explain why:

4E. Diabetes predisposition: under "disease predisposition," find "Type 1 diabetes" and "Type 2…"

Does Greg Mendel have lower than average risk of developing type 1 diabetes? Yes No

Does Greg Mendel have lower than average risk of developing type 2 diabetes? Yes No

How many SNPs were used to predict predisposition for type 2 diabetes? _____

Do all SNPs make similar predictions: all increased or all decreased risk of disease?

It is possible that there are other SNPs in the human genome that have not been discovered yet. Imagine,that Greg Mendel had his whole genome sequenced and knew **ALL** the SNPs affecting his predisposition to type 1 and type 2 diabetes.

For what type of diabetes would scientists/doctors be able to predict the patient's risk with more confidence? Type 1 Type 2 Explain.

If genetic information suggested lower than average risk of developing diabetes, could the patient still develop the disease? Yes No Explain

4F. What else can we learn about Greg Mendel from his genotype? Pick one trait that is of interest to you and summarize your findings.

Part 5. Complete Table 2. Summary of predicted phenotypes and identification.

Trait (phenotype)	X	Y	Z
Lactose intolerance	No	No	YES, lactose intolerant
Height (average, taller o shorter than average)			
Eye color (Blue, Brown, Green)			
Hair curl			
Longevity			
Name of the person			

Conclusion: Why is it difficult to predict traits from just knowing DNA letters at each SNP? (Can you think of three reasons?)