**H+, or how to build a perfect human.**

**Introduction**

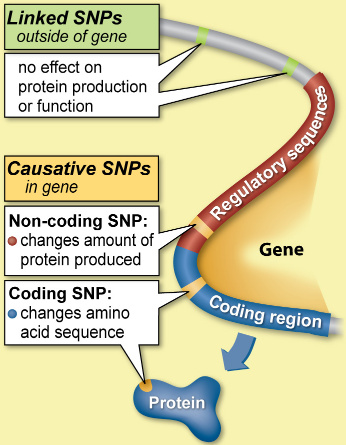
Each of us carries a large number of genetic variations. Nowadays, we have a lot of services that allow you to get information about your genome. It can be whole genome sequence obtained by an NGS instrument or just collection of SNPs obtained using a genotyping chips (like Illumina HumanOmniExpress-24 used by 23&Me that can test about 700k known SNPs).

We can use SNPs to predict likelihood of having some phenotypic trait, and, more important, likelihood of disease. And this information may require actions - changing in lifestyle, or even medical intervention.

SNPs are not necessarily located within genes, and they do not always affect the way a protein functions. We can divide SNPs into two main categories:

1) Linked, or indicative SNPs. They usually do not reside within genes and do not affect protein function. But they correspond to particular traits (like a certain drug response or risk for getting a certain disease).

2) Casuative SNPs, that affect the protein function. They can be located in the coding region, changing amino acid sequence, or within the regulatory sequence, changing level or timing of gene expression.

  
© Genetic Science Learning Center, University of Utah

Of course, there are lots of possible concerns. Someone might just not want to know (“ignorance is bliss”), in some cases there can be some ethical issues (such as unwanted adoption disclosure), sometimes information about you can reveal information about your relatives. Genetic data are intrinsically self-identifying, and even if the data sets are assumed to be anonymous after stripping off personal identifiers, re-identification is possible. Revealed genetic information possibly can affect your life in some cases: it’s not there yet, but we can easily imagine that predisposition to certain diseases can raise your insurance rates, or affect a decision to give you a job. Despite these privacy issues, some people are releasing their data as open source projects, believing that their genomes will help scientists gain insights about the human genome and that the benefits outweigh the drawbacks.

For this week’s project, let’s imagine that we are in the not-too-distant-future, where transhumanism has been widely accepted, and we are allowed to use CRISPR-Cas9 on humans. Imagine you can just order a DIY kit to make any corrections to your DNA (actually, you can [order it now](http://www.the-odin.com/diy-crispr-kit/), but just for *E. coli*). What would you change?

**Dataset**

We will work with raw 23andMe data. You can pick one of three datasets:

1. “GitHub Guy” (<https://github.com/msporny/dna>): 23andMe results released under Creative Commons Public Domain License.
2. “FixMyProfessor.com” - [my 23andMe data](https://drive.google.com/open?id=1QJkwJe5Xl_jSVpqdTSNXP7sqlYfI666j) (yes, I decided to release it for class purposes)
3. \*If you have your own 23andMe data, you can use it. Note that we will use the [23andMe privacy statement](https://www.23andme.com/about/privacy/) and no one except your instructor will have access to your reports.

For each detected SNP, you have its chromosome, position, and a unique identifier (rsid). Using this information, you can query different databases (e.g. dbSNP, ClinVar, SNPedia, OMIM) and get information about the SNP. For some steps of your analysis, you may want to turn your data into a vcf file - there are a lot of options for this task, e.g. bcftools convert or plink. NB: use the proper human genome reference version - for 23andMe results before August 9, 2012 it is GRCh36, [later - GRCh37](https://customercare.23andme.com/hc/en-us/articles/212883767-Which-reference-genome-does-23andMe-use-).

**Tl;dr:** Download [plink](https://www.cog-genomics.org/plink/) and try:  
plink --23file <input\_file.txt> --recode vcf --<output.vcf>

This command will get you vcf file in GRCh37 coordinates, so you may download [corresponding vcf from ClinVar database](https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/) and intersect:

bedtools intersect -a <your\_vcf> -b <clinvar\_vcf> > <intersected\_vcf>

keeping only clinically relevant SNPs. Also, you can just grep by SNP id - they are consistent along the versions of the genome.

This project is open-ended: there is no single correct way to do this, but for the sake of grading, you will need to provide the following:

1. “Who we are?”

1. Origin. Establish probable ethnicity and (optionally) maternal (mtDNA) and paternal (Y-chromosome) haplogroups.
2. Annotation. Use Variant Effect Predictor, ANNOVAR, SNPEff or any other tool of your choice to annotate the obtained file. If you succeed, you’ll immediately become a valuable specialist on the job market (in case you haven't forgotten how to make a vcf file from the raw reads).

2. “Where are we going?”

1. Fixes - specific changes you would make to fix some issues. Provide minimum five.
2. Improvements - specific changes you would make to give the person in question some advantages. Also provide minimum five.

You may consider Ensembl BioMart (with powerful query interface to most common databases and corresponding R/Python packages) or any other tools of your choice (Interpretome/GENOtation, Enlis Genome etc). For mtDNA <http://haplogrep.uibk.ac.at> can be handy. Also, you can download annotated vcf files from ClinVar or other databases, and intersect with your data.

There are plenty of blog posts considering playing with 23andMe output - for example, [this analysis using Google Genomics](https://medium.com/google-cloud/interpreting-23andme-raw-genome-data-with-google-genomics-bigquery-and-cloud-datalab-f8540b6b7ef9). Also, consider series of the posts ([one](https://medium.com/@german.m.demidov/how-to-analyse-your-own-dna-a-personal-experience-c4057d41753f), [two](https://medium.com/@german.m.demidov/how-to-analyse-your-own-dna-a-point-of-view-of-ordinary-customer-part-i-226284ba9466), [three](https://medium.com/@german.m.demidov/how-to-analyse-your-own-dna-a-point-of-view-of-ordinary-customer-part-ii-dc558557db36)) by German Demidov - along with the technical side of things, it gives you a feeling how it really looks like analysing your own genomic data.

Extra credit (1 point): provide guide RNAs for your findings.

**For your lab report:**

Intro: Provide biological motivation and some background on the technology behind the project (genotyping chips and CRISPR-Cas9).

Methods: Briefly explain the raw data, tools, and databases you used.

Results: Provide information you were able to pull out from the chosen dataset. List all the suggested changes.

Discussion: For suggested changes, describe the mechanism of action (if known) - what was changed on the protein or regulatory region level, and how exactly it affects the person.