# Resource pack:
## Human genetic variation and disease


### Scottish Curriculum Links

#### Higher Biology: DNA and the Genome Unit

Study of DNA and the genome. The Unit covers the key areas of structure of DNA; replication of DNA; control of gene expression; cellular differentiation; the structure of the genome; mutations; evolution; genomic sequencing; personal genomics.

#### Higher Human Biology: Human Cells Unit

Structure and Replication of DNA, Gene Expression, Genes and Proteins in Health and Disease, Human Genomics. The Unit covers the key areas of structure and replication of DNA; gene expression; genes and proteins in health and disease, human genomics. Expression of genotype and protein production, which allows study of mutations and disorders

#### National 5 Biology: Multicellular Organisms Unit

Reproduction, variation and inheritance; the need for transport and effects of life-style choices on animal transport and exchange systems.

Contents



For more information on obesity and diabetes, mentioned as example diseases here, another resource pack is available: www.sserc.org.uk/images/Biology/Higher_Human/SQA/Diabetes_final.pdf

## Background

Many common diseases are influenced by a combination of multiple genes and environmental factors and are therefore referred to as complex diseases. Because they can be caused by both genetic and environmental factors, complex diseases – such as asthma, stroke and diabetes - can be difficult to treat.

In 1990, a massive international scientific project called The Human Genome Project was launched to determine the sequence of the building blocks/letters/nucleotides of the human DNA code - the entire human genome. The project was a success and the data was made available in 2003.

Uncovering the sequence of the “average” or “reference” human genome was only the first step in understanding how the instructions coded in DNA provide the basis for all biological processes and can lead to disease.


Researchers, including many at the MRC Human Genetics Unit at the University of Edinburgh, are now looking at genetic variations between people to determine their importance and predict risk of particular diseases and response to certain medications.

In order to understand how genetic variation is related to disease it is necessary to have health and other information about a person (their phenotype), as well as their DNA. One approach is to study large groups of people, called cohorts. A cohort is a group of people with a shared characteristic, for example, their age (a birth cohort), geographical location, or the fact they suffer from a specific disease. Generation Scotland is an example of a cohort that was recruited in Scotland to help improve health by understanding disease. Cohort participants completed questionnaires about their health and lifestyle, underwent medical examinations and gave samples including DNA, blood and urine for use in medical research. These samples are stored in a biobank and can be accessed by researchers to answer their own research questions.

As genomic technology continues to improve and costs decrease, it is anticipated that health professionals will be able to provide patients with personalised information about their risks of developing certain diseases, tailor prevention programs to each person's unique genetic makeup and select the treatments most likely to be effective and least likely to cause adverse reactions in that particular patient.

## Human genetic variation

Any two humans are, genetically, around 99.5 per cent the same.

The most common type of genetic variation among people is known as a single nucleotide polymorphism or SNP (pronounced “snip”). SNPs are single base pair variations in DNA sequences which exist normally throughout a person’s DNA and most have no effect on health or development. Some of these genetic differences, however, have proven to be very important for human health.

Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may represent the replacement of the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain position of DNA in the reference human genome. This is known as a genetic variant. It might also be described as a mutation but the term variant is used by geneticists as the effect of having either nucleotide might be completely neutral.

Most commonly, SNPs are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease or with traits that may influence the risk of disease, for example high blood cholesterol increases the risk of cardiovascular disease. A person’s genotype refers to the genetic variation they carry across the genome and can be relevant for a single trait or set of traits.

When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene’s function by altering protein sequence or by changing the amount of expression of a gene, which results in different protein levels.

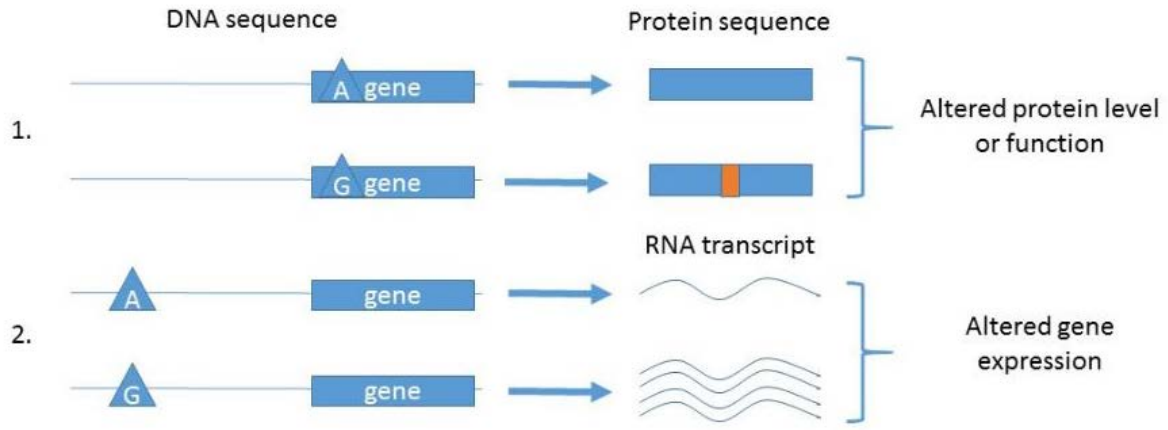
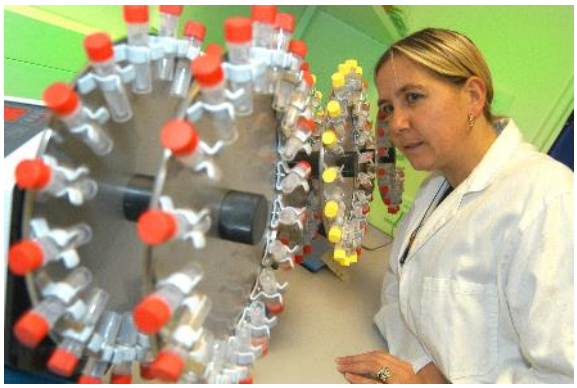



*Figure 1: DNA sequence variation may lead to 1) changes to protein sequence, structure or function when within the gene coding region 2) changes to level or pattern of gene expression if in a non-coding, regulatory region. This can result in changes to protein level.*

## Biobanks: giving researchers access to genetic & health information

Researchers who wish to use genetics to learn more about risks of disease need to be able to compare information about the genotype and phenotype of different individuals. In order to do this researchers need access to information from real people.

This often happens through cohort biobanking studies, like Generation Scotland, where people consent to give their DNA and other samples, complete carefully designed questionnaires about their lifestyle and medical history and undergo physical examination and blood tests.



Data, including SNP genotype frequencies and quantitative trait values, across populations such as Generation Scotland help researchers at the MRC Human Genetics Unit to conduct genome-wide association studies (GWAS) using many hundreds of thousands of SNPs. Genetic association tests can then be performed, using a range of statistical techniques and powerful computing clusters to understand the risks of developing diseases and ultimately lead to new treatments.

## Genome-wide association studies (GWAS) to predict disease risk

Genome-wide association studies (GWAS) are used by researchers to look at genetic variation/SNPs between people and predict disease risk.

SNPs can act as biological markers, helping scientists locate genes that are associated with a disease or with traits that may influence the risk of disease.

To locate genes that are associated with particular diseases, researchers can genotype two groups of participants using GWAS:

- people with the disease being studied (cases)
- similar people without the disease (controls).

If certain SNPs are found to be significantly more frequent in people with the disease compared to people without disease, the genetic variants are said to be "associated" with the disease. The associated genetic variations can serve as powerful pointers to the region of the human genome where the disease-causing problem resides. However, the associated variants themselves may not directly cause the disease. They may just be "tagging along" near to the actual variants that cause the biological change (see Figure 1). For this reason, researchers usually need to take additional steps, such as sequencing DNA base pairs in that particular region of the genome, to identify the exact genetic change involved in the disease.

GWAS can also be used to find associations between a trait that may influence the risk of disease and genotype:

1. Choose a group of participants and measure a trait e.g. weight, blood cholesterol, height
2. Genotype the participants SNPs
3. Find which SNPs tend to correlate with your trait.

*Figure 2* shows an example of this.

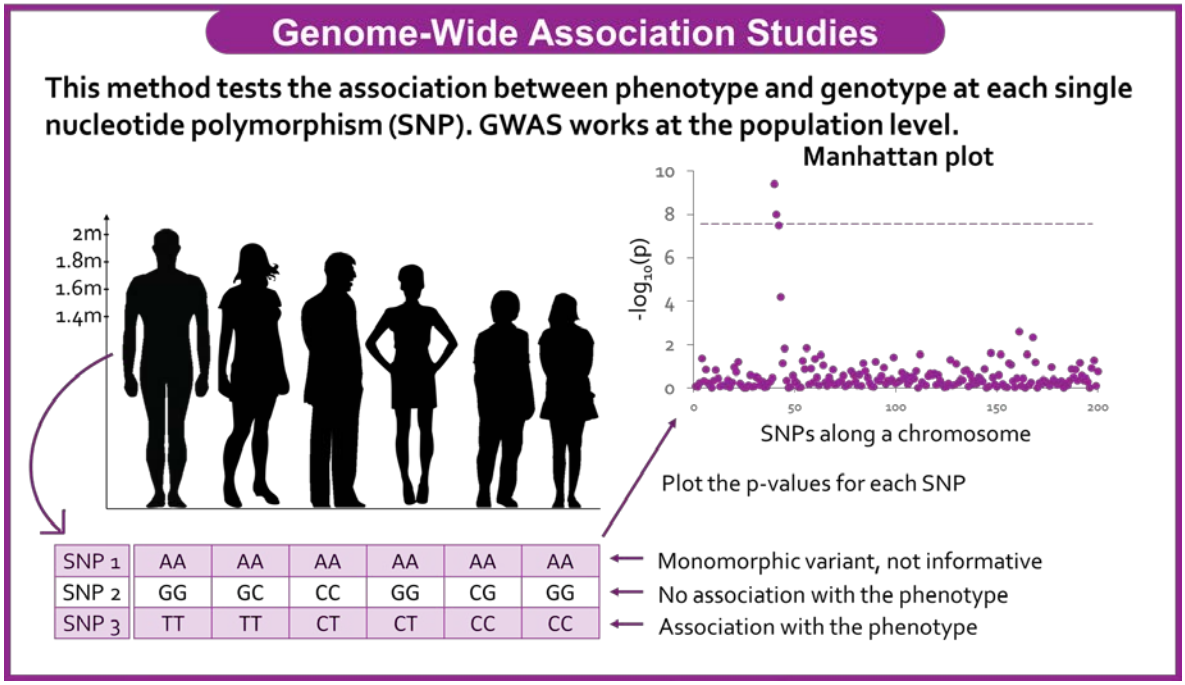



*Figure 2: Genome wide association study (GWAS) testing the association between height and genotype. For the first SNP, we see that everyone has the same allele, so this is not informative. In the second SNP, we see that both tall people and short people have the G allele, so this does not correlate with our trait. The third SNP, however, shows that taller people tend to have more T alleles while shorter people tend to have more C alleles, so we might have a good candidate here. When you do this for thousands of SNPs and plot the log of the p-value (see Statistical terms primer) on a graph, you end up with a 'Manhattan plot' that highlights significant regions.*

## Discovering information on specific genes and mutations identified by GWAS

Given the large amount of genetic and biological data available, it is important that specific resources exist to help researchers find out more about the SNPs and genes identified in their results. The US National Institutes of Health (NIH) have developed many resources to meet this need.

**Information about SNPs:**
The NIH dbSNP resource provides information for researchers who have identified specific SNPs in their studies. By entering the SNP ID (which begins with rs, or ss) the researcher can find out information about the SNP, for example its chromosomal location, sequence change represented, how common the alleles are in the population, whether it is in a gene, and links to normal (ancestral) sequence.

**Information about genes linked to diseases:**
The **NIH Genetics Home Reference resource** contains a wealth of information on genes and genetics. It can be used by researchers, once they identify a DNA change that may affect a specific gene, to find out whether that gene has already been linked to a disease.

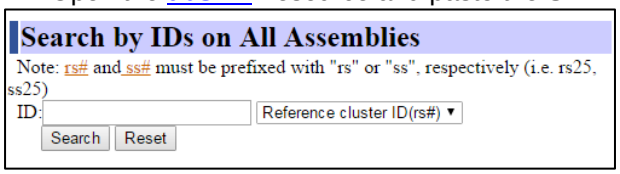
**Information about genes not yet linked to a disease:**
The NIH Gene database provides a comprehensive list of genes, where function has not yet been linked to a specific disease.

Once a gene has been identified, it is considered a 'candidate' gene for the disease or trait. Further study is needed to link genetic changes to disease. Researchers may choose to use model organisms, such as zebrafish or mice, which can be genetically manipulated to assess whether the gene variant has an effect on a relevant physical characteristic, or disease risk.

**Example 1**

A group of scientists performed genetic analysis to find SNPs associated with the risk of colon cancer. They identified the following statistically significant SNPs:
rs1321311, rs3824999, rs5934683

- Open the dbSNP resource and paste the SNP ID into the 'ID' field and search




- REF SNP alleles will tell you the two possible DNA bases
- Ancestral alleles will tell you which one is the usual allele
- Minor allele frequency tells you how common the rare allele is (this could help in estimating how many people in a population are likely to be heterozygous or homozygous for either allele)
- Clicking on 'chromosome position ('Chr Pos') to see if any genes are in the region

**Example 2**

One of the SNPs identified in the study above is near the POLD3 gene. This change might change the function or level of expression of the gene. To learn more about the gene and whether it might be related to colorectal cancer:

- Enter the NCBI gene database (http://www.ncbi.nlm.nih.gov/gene/)
- Search for POLD3
- In the search results ensure you select the human POLD3

## Real research data linking genetics, obesity and Type II diabetes

One specialism of the MRC Human Genetics Unit is the study of quantitative trait loci. A quantitative trait is any feature of a person that can be measured as a continuous variable. Examples of quantitative traits include height, weight, cholesterol, abdominal fat, body mass index (BMI) and lung capacity. Quantitative traits underlie biological differences between people and can affect health and disease risk.

A quantitative trait locus (QTL) is a section of DNA (the locus, plural loci) that correlates with continuous variation in a phenotype (the quantitative trait). The QTL typically is linked to, or contains, the genes that influence that phenotype. QTL analysis is particularly useful to study complex traits such as obesity and diabetes. One challenge for QTL studies is that traits can be affected by a combination of factors including genotype, environmental factors and interaction of these with specific genotypes, and even interactions between different genes.

The resulting dataset can be examined to find trends and correlations between the genotype and factors that can influence disease. A definition of the statistics terms relevant to these studies is included in this resource pack. For example, researchers could assess whether certain genotypes increase the risk of someone having high cholesterol, or obesity. The researchers do not have a specific hypothesis as to which genetic variants contribute to the disease. Instead they examine all of the genetic variation in the genome and use statistics to determine which genotypes are most associated with a specific condition, or trait. Entire new avenues of research can then be opened up, which may reveal biological processes important to disease, some of which could be targeted with existing or new drugs. The information could also be used to advise people on their risk of disease and whether specific preventive action would be helpful.

*Table 1: Individual Level Data from Participants in Generation Scotland (Scottish Family Health Study, GS:SFHS)*

| Family ID | Volunteer ID | father ID | mother ID | sex | Age (years) | In study | Genotype (rs1421085) | height (cm) | weight (kg) | waist (cm) | hips (cm) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 126036 | 0 | 0 | F | 66 | Y | TT | 163 | 87.2 | 95 | 115 |
| 1 | 160864 | 0 | 0 | M | | N | | | | | |
| 1 | 106845 | 160864 | 126036 | F | 46 | Y | TT | 162 | 85.6 | 100 | 110.5 |
| 1 | 8627 | 160864 | 126036 | F | 43 | Y | CT | 160 | 60.8 | 70 | 102 |
| 1 | 135861 | 160864 | 126036 | F | 40 | Y | CT | 159 | 73.2 | 75 | 93 |
| 2 | 155447 | 0 | 0 | F | 51 | Y | TT | 165 | 57.9 | 82 | 99 |
| 2 | 87337 | 0 | 0 | M | 48 | Y | CC | 185 | 86.3 | 95.5 | 109 |
| 2 | 143760 | 87337 | 155447 | F | 28 | Y | CT | 178 | 61.2 | 69.5 | 99 |
| 2 | 131643 | 87337 | 155447 | M | 24 | Y | CT | 189 | 93.2 | 96 | 109 |
| 2 | 136541 | 87337 | 155447 | F | 22 | Y | CT | 164 | 63.4 | 76 | 104 |
| 3 | 18208 | 0 | 0 | F | 68 | Y | TT | 155 | 82.4 | 92 | 102 |
| 3 | 30403 | 0 | 0 | M | 69 | Y | TT | 168 | 81.3 | 102 | 106 |
| 3 | 7241 | 0 | 0 | F | | N | | | | | |
| 3 | 160691 | 0 | 0 | M | | N | | | | | |
| 3 | 134197 | 8605 | 139455 | M | 21 | Y | CT | 171 | 69.8 | 88 | 102.5 |
| 3 | 24176 | 8605 | 139455 | F | 19 | Y | CT | 160 | 57.9 | 77 | 98 |
| 3 | 139455 | 30403 | 18208 | F | 44 | Y | TT | 161 | 67.9 | 87 | 101.5 |
| 3 | 142039 | 30403 | 18208 | F | | N | | | | | |
| 3 | 114684 | 160691 | 7241 | M | 53 | Y | CT | 160 | 81.3 | 102 | 107 |
| 3 | 8605 | 160691 | 7241 | M | 49 | Y | CC | 175 | 91.2 | 103 | 109 |

*Table 1 shows the volunteer identity number (ID), identity numbers of their mother and father if in the study, sex (male or female), age and genotype (CC, CT or TT) for a SNP (rs1421085) in three families (Family IDs 1, 2 and 3) in GS:SFHS. So that private information is protected and researchers do not know who each study volunteer is, ID numbers are assigned. Entries of "0" indicate the data is not available. Measurements of height (cm), weight (kg), waist circumference (cm) and hip circumference (cm) are given. (http://dx.doi.org/10.7488/ds/1581)*

Other studies on the genetics of obesity and diabetes have been performed with the UK Biobank (www.ukbiobank.ac.uk). One study examined how genotype can affect the risk of being obese or severely obese, using the concept of an **odds ratio**.

Tables 2, 3 and 4 below show data from GWAS studies examining the association of SNPs with BMI, obesity and Type 2 diabetes in 119,688 people in the UK Biobank. Significant results were found at the FTO locus, for BMI and obesity, and at a locus in the *CDKAL1* gene for Type 2 Diabetes.

*Table 2: BMI values by genotype group at the FTO locus (rs1421085)*

| **rs1421085 genotype group** | **TT** | **CT** | **CC** |
|---|---|---|---|
| *Number of participants (n)* | 42,835 | 57,524 | 19,329 |
| Mean BMI kg/m$^2$ (95% Confidence Interval) | 27.27 (27.22, 27.31) | 27.54 (27.50, 27.58) | 28.07 (28.00, 28.14) |

*n* = 119,688 British individuals in the UK Biobank

At the FTO locus, the C allele is associated with higher BMI (Table 2) and it is therefore more likely that those with a copy of this allele will be obese. (Table 3). Those with two copies of the C allele are a little more than twice as likely to be severely obese, as you can see from the odds ratio of 2.12 when comparing people with the TT and CC genotypes.

*Table 3: Odds ratios for 'obese' and 'severely obese' classifications by genotype group at the FTO locus (rs1421085)*

| **Genotype** | **TT vs CT** | | **CT vs CC** | | **TT vs CC** | |
|---|---|---|---|---|---|---|
| | **Odds ratio (95% confidence interval)** | ***p* value** | **Odds ratio (95% confidence interval)** | ***p* value** | **Odds ratio (95% confidence interval)** | ***p* value** |
| **BMI Class** | | | | | | |
| Obese | 1.15 (1.12, 1.19) | $2\times10^{-16}$ | 1.28 (1.23, 1.34) | $3\times10^{-28}$ | 1.48 (1.41, 1.55) | $2\times10^{-62}$ |
| Severely obese | 1.28 (1.16, 1.41) | $7\times10^{-07}$ | 1.66 (1.49, 1.85) | $2\times10^{-20}$ | 2.12 (1.88, 2.38) | $2\times10^{-36}$ |

In Table 4 you can see evidence that, in the UK Biobank, a change from A to G in the *CDKAL1* gene increases the chance of developing Type 2 Diabetes.

*Table 4: Type 2 diabetes ORs by genotype group at a different locus in the CDKAL1 gene (rs7756992)*

| **AA vs AG** | | **AG vs GG** | | **AA vs GG** | |
|---|---|---|---|---|---|
| **Odds ratio (95% confidence interval)** | *p value* | **Odds ratio (95% confidence interval)** | *p value* | **Odds ratio (95% confidence interval)** | *p value* |
| 1.06 (0.99, 1.14) | 0.077 | 1.39 (1.24, 1.56) | $1 \times 10^{-8}$ | 1.48 (1.32, 1.65) | $6 \times 10^{-12}$ |

*n = 117,775 British individuals in the UK Biobank*

## Statistics terms

**Mean** – the mean of a sample is the average generated by adding all measured values together and dividing by the total number of measures.

**Median** – The median is the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. In an odd numbered data set this would be the middle value, in an even numbered set it would be the mean of the 2 middle values

**Range** - The "range" is the difference between the largest and smallest values and can be used to show the total variation in a population.

**Confidence interval (CI)** - this describes the uncertainty of a sampling method. A CI is expressed as a range of values and a percentage. For example, the mean BMI (95% confidence interval) of the C/C genotype in Table 2 is given as 28.07 (28.00,28.14). The first number is the mean. This means that if different samples from the same population were sampled in the same way, that the true mean of the sample would fall between these values 95% of the time. In a CI, the range of values above and below the sample statistic is called the margin of error.

**Odds Ratio (OR)** - a measure of association between an exposure, for example a genotype, other variable, and an outcome, such as a specific disease. The OR describes the odds of the outcome happening with compared to without the specific exposure.

For example, you could compare a specific genotype (A) to a different genotype (B) and see the effect on a disease outcome, for example diabetes.

- Odds ratio equal to 1 for B compared to A would suggest that the two genotypes do not have different influences on diabetes occurrence.
- Odds ratio < 1 (e.g. 0.5), for genotype B compared to A suggests that one genotype results in less occurrence (in this example 50%) of diabetes as the other genotype, which could mean a protective effect.
- An odds ratio of >1 (e.g. 3) for genotype B compared to A would suggest that there is a higher chance of diabetes with genotype B (in this example 300% or 3 times the chance), which could imply a detrimental effect of that genotype on health

When making these comparisons the statistical confidence is usually expressed with a p value.

**p value** - represents the probability that the observed result has nothing to do with what one is actually testing for and is generated using one of several statistical tests that must be carefully chosen depending on the hypothesis. The smaller the p value, the more likely that the observed result is not due to chance.

**n** – is used to denote the sample size. For example in Table 1 there are 2 volunteers with the CC genotype, i.e. $n$=2

## Glossary

**Biobank**: a biobank is a collection of biological samples (usually human) and associated data for use in research

**Biological marker:** also known as a biomarker, a characteristic that can be measured as an indicator of a biological process

**Body Mass Index (BMI)**: a measure that adults can use to find out if they are a healthy weight for their height.

**Cohort**: a group of people with a shared characteristic, for example, their age (a birth cohort), geographical location, or the fact they suffer from a specific disease

**Complex disease**: disease influenced by a combination of multiple genes and environmental factors

**Generation Scotland**: Partnership between the Scottish University Medical Schools, the NHS in Scotland and the people of Scotland. Over 30,000 people from across Scotland have helped Generation Scotland become a world class biomedical resource for research into a wide variety of diseases, such as heart disease, diabetes and mental health problems. They did this by participating in one of the research studies and contributing blood and other samples, clinical measurements such as blood pressure and information about their health and lifestyle. This research resource will support medical research in identifying the genetic basis of common complex diseases.

**Genetic variation**: the variation in the DNA sequence in each of our genomes. Genetic variation is what makes us all unique

**Genetic variant**: one of multiple alternative forms within a nucleic acid sequence

**Genome-wide association studies (GWAS)**: examine many common genetic variants in different individuals to see if any variant is associated with a trait

**Genotype**: the genetic variation across the genome

**Human genome**: the sequence of the building blocks/letters/nucleotides of the human DNA code

**Locus:** A locus (plural loci), in genetics, is the specific location or position of a gene's DNA sequence, on a chromosome.

**Major allele**: the sequence variant of a specific piece of DNA that is the most common in the population

**Odds ratio**: a measure of association between an exposure and an outcome. The OR describes the odds of the outcome happening with compared to without the specific exposure

**Phenotype**: health and other measurable information about a person

**Single nucleotide polymorphism or SNP** (pronounced “snip”): single base pair variations in DNA sequences which exist normally throughout a person’s DNA and most have no effect on health or development. Some of these genetic differences, however, have proven to be very important for human health.

# Sources/further reading

## Genetics and DNA variation

Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Dunlop MG *et al*. Nature Genetics 2012 May 27;44(7):770-6. doi: 10.1038/ng.2293.

Your Genome http://www.yourgenome.org/facts

NIH National Human Genome Research Institute, USA
www.genome.gov/10000202/fact-sheets

dbSNP database of SNPs http://www.ncbi.nlm.nih.gov/SNP

NIH US National Library of Medicine "Help me understand genetics"
https://ghr.nlm.nih.gov/primer

NCBI gene database http://www.ncbi.nlm.nih.gov/gene/

www.dnadarwin.org DNA to Darwin allows 16–19 year-old school students to explore the molecular evidence for evolution through practical bioinformatics activities that use data analysis tools and molecular data

Animal models for human disease
http://www.nature.com/scitable/topicpage/the-use-of-animal-models-in-studying-855

## Biobanks and the genetics of complex disease

Generation Scotland
http://www.ed.ac.uk/generation-scotland

UK Biobank
http://www.ukbiobank.ac.uk

Body Mass Index NHS primer
http://www.nhs.uk/chq/Pages/how-can-i-work-out-my-bmi.aspx?CategoryID=51

A study showing interaction between FTO genotype and BMI
www.ncbi.nlm.nih.gov/pubmed/22982992

FTO genotype and diabetes UK Biobank Study
http://link.springer.com/article/10.1007%2Fs00125-016-3908-5

Interaction between FTO genotype and BMI - additional study
www.ncbi.nlm.nih.gov/pubmed/22982992

Biointeractive- mapping genes to traits in dogs
www.hhmi.org/biointeractive/mapping-genes-traits-dogs-using-snps

A beginner's guide to interpreting odds ratios, confidence intervals and p values
http://www.students4bestevidence.net/a-beginners-guide-to-interpreting-odds-ratios-confidence-intervals-and-p-values-the-nuts-and-bolts-20-minute-tutorial/

Diabetes risk calculator
http://www.nhs.uk/Tools/Pages/Diabetes.aspx

## Teacher's information pack

Teachers can obtain an information pack relating to this resource by contacting the Lead Biology Teacher in their Education Authority.maybe change to region to Education Authority

## Acknowledgements/terms of use

Table 1 in this resource is from *Family Genotype and Phenotype Data* [dataset]. MRC Institute of Genetics & Molecular Medicine at the University of Edinburgh, MRC Human Genetics Unit, QTL Collection. Kerr, Shona; Campbell, Archie; Porteous, David; Hayward, Caroline. (2016). http://dx.doi.org/10.7488/ds/1581 used under CC BY 4.0.

Tables 2, 3 and 4 in this resource are derivatives of Tables 2, 3 and 5 in *Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively*. Wood AR et al. Diabetologia. 2016 Jun;59(6):1214-21. doi: 10.1007/s00125-016-3908-5 used under CC BY 4.0.

## Referencing this resource

Please use "Human genetic variation and disease resource pack", http://www.ed.ac.uk/mrc-human-genetics-unit/public-events-resources/inspiring-the-next-generation-of-researchers/school-data-resources