Human genetic variation (I): an introduction

Genetic variation is fundamental to the evolution of all species and is what makes us individuals. Our genes have a large influence on our lives. They affect what we look like, our personalities and preferences and our susceptibility to disease. By studying genetic variation we hope to understand the molecular process that contribute to life on earth.

The study of genetic variation has been used to model human migration, understand the cause of human diseases, and to predict disease outcomes.

This is part I of our course on human genetic variation. In this part of the course we introduce some key concepts in the field of human genetic variation including the types of variants, variant effects and genetic variation studies.

In part II of this course [9] we will learn how to explore publicly available genetic variation data.

The courses focus on heritable (germline) variation and will give you a taste of the resources you can use to explore genetic variation data.

Learning objectives:

- Review sources of genetic variation
- Describe the possible effects of genetic variation
- Identify common genetic variation file types and formats
- Describe types of genetic variation studies
What is genetic variation?

Origins of genetic variation

Genetic variation is the difference in DNA sequences between individuals within a population. Variation occurs in germ cells i.e. sperm and egg, and also in somatic (all other) cells. Only variation that arises in germ cells can be inherited from one individual to another and so affect population dynamics, and ultimately evolution. Mutations and recombination are major sources of variation.

Mutations are the original source of genetic variation. A mutation is a permanent alteration to a DNA sequence. De novo (new) mutations occur when there is an error during DNA replication that is not corrected by DNA repair enzymes. It is only once the error is copied by DNA replication, and fixed in the DNA that it is considered to be a mutation (Figure 1). Mutations may be beneficial to the organism; deleterious (harmful) to the organism; or neutral (have no effect on the fitness of the organism).

Somatic mutations can accumulate in our cells and are mostly harmless. They can lead to local changes in tissues such as moles appearing on the skin, and can also have more serious effects - for example leading to cancer. To learn more about the role of somatic mutations in cancer have a look at this paper by Martincorena and Campbell [10] [11]. In this course we focus on heritable genetic variation, i.e. variation that occurs in germ cells.

Recombination is another major source of genetic variation. Each of us has a mixture of genetic material from our parents. The mixing of this genetic material occurs during recombination when homologous DNA strands align and cross over. Recombination effectively ‘shuffles’ maternal and paternal DNA, creating new combinations of variants in the daughter germ-cells (Figure 2).


Recombination is another major source of genetic variation.
Figure 2 Recombination contributes to human genetic variation by shuffling parental DNA and creating new combinations of variants. Image source: Creation Wiki [14].

For a discussion on the sources of variation within a population see this chapter in An Introduction to Genetic Analysis. 7th edition [15] [16].

Variants, alleles and haplotypes

In the field of genetic variation, the term **variant** is used to refer to a specific region of the genome which differs between two genomes.

Different versions of the same variant are called **alleles**. For example, a SNP may have two alternative bases, or alleles, C and T [17].

When working with genome scale data the term **reference allele** refers to the base that is found in the reference genome. Since the reference is just somebody’s genome, it is not always the major allele. In contrast, the **alternative allele** refers to any base, other than the reference, that is found at that locus. The alternative allele is not necessarily the minor allele and it may, or may not, be linked to a phenotype. There can be more than one alternative allele per variant.

In the genome, alleles at variants close together on the same chromosome tend to occur together more often than is expected by chance. These blocks of alleles are called haplotypes. **Linkage disequilibrium** (LD) is a measure of how often two alleles or specific sequences are inherited together, with alleles that are always co-inherited said to be in linkage disequilibrium.

Genotypes and phenotypes

**Genotypes** are the genetic make-up of an individual [18]. **Phenotypes** are the physical traits and
characteristics of an individual and are influenced by their genotype and the environment.[19].

Genetic differences or variation between individuals leads to differences in an individual’s phenotype, trait or risk of developing a disease. An individual's phenotype is influenced both by their genotype and their environment.

A mendelian trait is one that is controlled by a single locus, for example a single SNP. In most cases, associations between genetic variants and phenotypes/traits are not this simple. These are called complex phenotypes and may be influenced by multiple variants in the genome along with environmental factors.

**Types of genetic variation**

Genetic variation is commonly divided into three main forms:

**Single base-pair substitution, also known as single nucleotide polymorphism (SNP)**

Any nucleic acid substitution, whether this is a transition (interchange of the purine (Adenine/Guanine) or pyrimidine (Cytosine/Thymine) nucleic acids); or a transversion (interchange of a purine and pyrimidine nucleic acid) (Figure 3).

![Figure 3](image)

**SNP**

Reference: ACTGACGCATGCATCATGCATGC

SNP: ACTGACGCATGCATCATGCATCATTCATGC

**Figure 3** SNPs result from the substitution of a single base-pair. In this example we have a transversion event substituting a Thymine nucleic acid in place of a Guanine.

**Insertion or deletion, also known as ‘indel’**

Insertion or deletion of a single stretch of DNA sequence that can range from two to hundreds of base-pairs in length (Figure 4).

![Figure 4](image)

**Reference**

ACTGACGCATGCATCATGCATGC

**Insertion**

ACTGACGCATGCTACATCATGCATGC

**Deletion**

ACTGACG—TGACATCATGCATGC

**Figure 4** Indels affect a string of base-pairs. In this example the insertion string shows that GTA has been inserted, and the deletion string shows a deletion of CA.

**Structural variation**

Typically used to describe genetic variation that occurs over a larger DNA sequence. This category of genetic variation includes both copy number variation and chromosomal rearrangement events. The five most common types of structural variants are shown in Figure 5.
**Variants in coding regions**

If a variant falls within a coding region, it can be categorised based on how it would affect the codon it falls within (Figure 6).

Due to redundancies in the genetic code [20], many nucleotide changes will not change the amino acid [21] sequence, for example a GCT to GCC change would still encode an alanine. These changes are known as synonymous.

Other changes will turn a coding codon, such as GGA glycine, to a stop codon, e.g. TGA, and are known as nonsense. This will result in a truncated protein, which may or may not be subject to nonsense-mediated decay [22] depending on where in the peptide it occurs.

If the change results in a change in amino acid, for example ACC threonine to AAC asparagine, this change is called missense. Algorithms such as SIFT [23] and PolyPhen [24] estimate how likely this amino acid change is to affect protein function. These estimates are based on how well conserved the protein is, the chemical difference between the amino acids, and the 3D structure of the protein (PolyPhen only). Both provide a score out of one (0 is the most severe for SIFT, whereas 1 is the most severe for PolyPhen) along with a qualitative prediction. These are predictions only, not experimental validations of the effect.
Figure 6 Mutations are classified based on how they affect the codon they are in. Image source: Wikimedia commons [25].

Indels with a length divisible by three (i.e. whole codon indels) in coding regions will cause insertions or deletions of whole amino acids into the protein, and are known as **in-frame deletions** or **insertions**. Note that indels divisible by three may also cause a missense or nonsense variant if the the variant falls across two codons. However, if the length is not divisible by three, this will cause a **frameshift** where all codons downstream of the indel are shifted, often resulting in a malformed protein or nonsense-mediated decay.

Variants in transcription factor binding motifs

**Transcription factor binding motifs** [26] (TFBMs) are genomic sequences that specifically bind to transcription factors. The consensus sequence of a TFBM is variable, and there are a number of possible bases at certain positions in the motif, whereas other positions have a fixed base. These are usually illustrated in sequence logo diagrams (Figure 7), where the height of the letter represents how frequently that nucleotide is observed in that position. Known TFBMs are curated in the JASPAR [27] database.
Because TFBSMs are variable, all motifs found in genomes are given a score out of one, indicating how strong the TFBSM is. The score represents the probability of each base occurring at each location in the motif. If a variant hits a TFBSM, it will alter the motif score, making it more or less likely to bind the transcription factor. These variants are called TF binding site variants, and the change in the motif score can be calculated.

**Variant effects on protein structure**

The effects of variants on protein structure can vary dramatically depending on the type of protein and the extent of variation. Take, for instance, a large, multi-domain [30] protein with multiple variants. One variant could involve the deletion of a large region of protein sequence, corresponding to an individual domain. If the core function of these domains are independent, the function of the remaining domains may not be affected at all. Another variant of this protein, however, may involve the changing of only a single amino acid [21] in the protein sequence. Despite the relative insignificance of this change to the overall protein sequence, if this residue is key to the function of the protein (for instance at the active site [31] of an enzyme) then this could completely eradicate protein function, or modify it in some way, for instance to change the specificity of an enzyme. For further reading on this, please see this paper by Studer RA et al [32][33].

**Variants affecting core protein structures**

An important factor in the effect of variation on protein structure is in the packing of the amino acids in the core of the protein structure. Most protein side chains are designated as either hydrophilic residues or hydrophobic residues. For a soluble protein, the outer surface of the protein is mainly composed of hydrophilic residues, whereas the core of the protein involves the packing of hydrophobic residues, away from the water in the solution. A variant that involves a change in the protein core from a hydrophobic residue to a hydrophilic one may destabilise the protein enough for function to be lost. Because the protein core is tightly packed, a change from a small amino acid [21] (e.g. alanine) to a bulkier one (e.g. phenylalanine) may not be accommodated and may lead to the protein being unable to fold into its 3D form.

An example of hydrophobic to hydrophilic variation is in human cystatin where a single variant, Leucine (L) to Glutamine (Q), destabilises the core of the protein. This variant, found to be more predominant in the Icelandic population, causes a large scale structural change leading to amyloid
formation and cerebral haemorrhage (see Figure 8 and PDBe's Quips article about this protein [34]).

Figure 8 In human cystatin a single variant causes Leucine (L) (purple) to be converted to a Glutamine (Q) in the variant structure, destabilising the core structure.

**Variants affecting surface residues**

Change of surface residues on a protein may affect its association with other proteins, though in practice changes in these regions have a less significant functional effect than changes in the core of the protein [35]. Should a protein be involved in protein-protein, or protein-nucleic acid interactions, any variation of amino acids on the binding surface could lead to a loss of function. An example of this is Human DJ-1, which in rare forms of Parkinson's disease, contains single mutations that destabilise the homodimeric interface, leading to disruption of function (Figure 9) [36].
Figure 9  **Structure of Human DJ-1** [37] from PDBe. Displayed in spheres is Methionine26. Mutation of Methionine26 to Isoleucine leads to destabilisation of the homodimeric interface in some forms of Parkinson's disease.

**Variation in prokaryotes**

In this course we focus on variation in humans who are sexually reproducing diploid eukaryotes. It is important to realise that there are several significant differences that must be considered if you are working on prokaryotes.
Figure 10 There are additional factors to consider when studying genetic variation in prokaryotes such as *E. coli* (pictured). Image source: Wikimedia Commons [38].

1. Prokaryotes tend to have haploid genomes, meaning that they only have one copy of each gene per individual. This affects the effective population size by comparison with a diploid species, where there are effectively twice the number of potential genetic variants for the same number of individuals.
2. Prokaryotes reproduce clonally (i.e. by making an exact copy of the cell), so recombination does not occur as a matter of course at every generation. Prokaryotes can also exchange genetic material laterally via transformation, conjugation and transduction.
3. Prokaryotes typically reproduce much more rapidly than eukaryotes, and these short generation times can lead to more rapid adaptation. This is why scientists are able to use certain prokaryotes to study evolution in the lab.

Have a look at this course [39] from Penn State University for an overview the cellular and genetic organisation of prokaryotes [40].

**Variant identification and analysis**

A likely workflow in human genetic variation studies is the analysis and identification of variants associated with a specific trait or population. Bioinformatics is key to each stage of this process and is essential for handling genome-scale data. It also provides us with a standardised framework to describe variants. In this section we will learn about the major steps in the process of variant calling [41], the VCF file format and variant identifiers. We will also examine the value of prediction in determining impact of variation on protein function and structure.

**What is variant calling?**

Variant calling is the process by which we identify variants from sequence data (Figure 11).

For whole genome or exome variant calling we follow a three step process:
1. Carry out whole genome or whole exome sequencing to create FASTQ \textsuperscript{42} files.
2. Align the sequences to a reference genome, creating BAM \textsuperscript{43} or CRAM \textsuperscript{44} files.
3. Identify where the aligned reads differ from the reference genome and write to a VCF \textsuperscript{45} file.

![A CRAM file aligned to a reference genomic region as visualised in Ensembl. Differences are highlighted in red in the reads, and will be called as variants.](image)

### Somatic versus germline variant calling

In germline variant calling, the reference genome is the standard for the species of interest. This allows us to identify genotypes. As most genomes are diploid, we expect to see that at any given locus, either all reads have the same base, indicating homozygosity, or approximately half of all reads have one base and half have another, indicating heterozygosity. An exception to this would be the sex chromosomes in male mammals.

In somatic variant calling, the reference is a related tissue from the same individual. Here, we expect to see mosaicism between cells.

### Understanding VCF format

VCF is the standard file format for storing variation data. It is used by large scale variant mapping projects such as IGSR \textsuperscript{46} and ExAC \textsuperscript{47}. It is also the standard output of variant calling software such as GATK \textsuperscript{48} and the standard input for variant analysis tools such as the VEP \textsuperscript{49} or for variation archives like EVA \textsuperscript{50}.

VCF is a preferred format because it is unambiguous, scalable and flexible, allowing extra information to be added to the info field. Many millions of variants can be stored in a single VCF file.

VCF files are tab delimited text files. Here is an example of a variant in VCF (Figure 12) as viewed in a spreadsheet:

![An example of a VCF file.](image)

### Table 1

<table>
<thead>
<tr>
<th>Column</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>#CHROM</td>
<td>Chromosome</td>
</tr>
<tr>
<td>POS</td>
<td>Co-ordinate</td>
</tr>
</tbody>
</table>

Each column in a VCF file provides a different type of information:
Human genetic variation (I): an introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

To learn more, take a look at the VCF specifications [45].

Variant identifiers

As we saw on the previous page, VCF files include variant identifiers. Identifiers are unique combinations of letters and numbers that are assigned to known entities, such as genes, variants and proteins. Variant identifiers are particularly useful when searching for information about a variant because they are unambiguous, unique and stable, unlike descriptive names, which can be used differently by different people, be identical between species and change over time.

Variants may have identifiers from multiple databases. You will see these different types of identifiers used throughout the literature and in other databases.

Different types of identifiers are used for short variants (Table 2) and for structural variants (Table 3). Some common databases and examples of the identifiers they use are shown in the tables below.

Table 2 Databases and identifiers for short variants (SNPs and indels).

<table>
<thead>
<tr>
<th>Identifier type</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssID [51]</td>
<td>ss335</td>
<td>Submitted SNP ID assigned by dbSNP or EVA</td>
</tr>
<tr>
<td>rsID [51]</td>
<td>rs334</td>
<td>Reference SNP ID assigned by dbSNP or EVA.</td>
</tr>
<tr>
<td>HGVS [52]*</td>
<td>ENST0000036667.4:c.803T&gt;C</td>
<td>Expresses the location of the variant in terms of a transcript or protein.</td>
</tr>
<tr>
<td>COSMIC ID [53]</td>
<td>COSM1290</td>
<td>ID assigned by COSMIC for somatic variants.</td>
</tr>
<tr>
<td>HGMD [54]</td>
<td>CD830010</td>
<td>ID assigned by HGMD to variants known to be deleterious.</td>
</tr>
</tbody>
</table>
*HGVS can be an ambiguous way to represent variants, so it is important to understand the format and its limitations. HGVS provide detailed documentation [52] on using this notation, including for indels, intronic variants etc.

**Table 3** Databases and identifiers for structural variants.

<table>
<thead>
<tr>
<th>Identifier type</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGVa [57]</td>
<td>essv8691751</td>
<td>Submitted structural variant ID assigned by DGVa. Variants are shared with dbVar.</td>
</tr>
<tr>
<td>dbVar [58]</td>
<td>nssv1602417</td>
<td>Submitted structural variant ID assigned by dbVar. Variants are shared with DGVa.</td>
</tr>
<tr>
<td>DGVa [57]</td>
<td>esv3364878</td>
<td>Variant region variant ID assigned by DGVa. Overlapping submitted variants (essv and nssv) are combined into a single variant region. The boundaries of a variant region may not match those of the submitted variants, which can vary.</td>
</tr>
<tr>
<td>dbVar [58]</td>
<td>nsv916030</td>
<td>Variant region variant ID assigned by dbVar. Overlapping submitted variants (essv and nssv) are combined into a single variant region. The boundaries of a variant region may not match those of the submitted variants, which can vary.</td>
</tr>
</tbody>
</table>

**Variant analysis**

Variants can be analysed in different ways. For example, you might want to determine which genes the variants hit and what effects they have on them. Tools such as the Ensembl VEP [49] and Snpeff [59] can be used for this. You may wish to filter or merge your VCF file, which you can do with VCF Tools [60].

Have a look at our webinar [61] to learn more about Ensembl VEP.

Having filtered your data and identified interesting genes or variants, you can explore them further using databases such as:

- Ensembl [62]
- Uniprot [63]
- GWAS Catalog [64]
- European Variation Archive (EVA) [50]

We will see some examples of this in the case studies [65] in part II of this course.

You may also wish to explore the variant consequences and any associated predictions. This can be useful for understanding why a particular variant may cause a particular phenotype. When studying a number of variants to determine which is causal for a phenotype, predictions can be helpful for prioritising variants for further study.
However, prediction can never be as valuable as experimental data. Any predictions determined bioinformatically should be followed up experimentally, and no conclusions should be drawn in what variant causes a phenotype based only on bioinformatic predictions.

For example, we can predict which amino acid changes are likely to affect protein function, but since we don’t fully understand how protein folding occurs, we cannot tell you if and how protein folding may change.

Now let’s take a quick look at some ways of predicting and visualising the effect of variation on protein structure and function.

### Predicting the effect of variation on protein structure and function.

The effect of genetic variation on protein structure and function varies dramatically depending on the type of protein and the extent of variation. This means that it is difficult to predict the exact effect of sequence variance upon structure, and therefore, function of a protein. The location of the variation needs to be considered, along with the function of the protein, to help gain an understanding of the effect of a variant on the protein structure and its function. There are various tools available to assist in the prediction of variation on protein stability, including:

- **DUET** [67]
- **I-Mutant** [68]
- **SDM** [69]

In addition, if there are structural homologs available, you can use structure prediction servers to predict the structure, for example **I-TASSER** [70] or **Phyre2** [71].

You should be aware, however, when using these tools that they are a prediction only and you should be careful not to draw too many conclusions based on these structural models.

### Visualisation of protein structure

If the protein structure is known, you can find the structure at the [Protein Data Bank](https://www.rcsb.org/pdb/) (PDB). Visualisation of the structure is possible with a number of downloadable programs, for example **PyMol** [73] and Coot, or on an online viewer, for example the LiteMol viewer used at PDBe.org.
Video below).

**Video 1** visualisation of the structure of haemoglobin using Litemol viewer.

With these viewers, you can view the location of a variant residue within the structure and use this information to assess whether this is likely to cause a significant effect in the protein function. At PDBe.org [72], on the macromolecules pages for an entry, you can access the Uniprot [74] sequence coverage viewer to find all other entries in the PDB with proteins from that Uniprot ID. This will then show, for each PDB entry, the regions of the sequence for which the structure has been solved. This will also highlight any variants in the structure, as compared to the UniProt sequence. If you would like to know more about how to find protein structures of interest at the PDBe then please watch our webinar PDBe: Searching for biological macromolecular structures [75].

**Types of genetic variation studies**

There are many ways in which you can study genetic variation however most studies can be loosely classified as:

- Genome wide association studies
- Studies on the functional consequences of variants
- Population genetics

In this section we highlight the key features of these types of studies and outline the ways that they have been used to answer biological questions.

**Genome wide association studies**

Many methods for associating variants with a phenotype, trait or disease rely on the fact that a variant leading to a phenotype is found at a higher frequency in cases (individuals with the phenotype) than controls (individuals without the phenotype). Genome wide association studies (GWAS) involve genotyping individuals at common variants across the genome using genome wide SNP [76] arrays. Variants associated with trait, or within the same haplotype [77] as a variant associated with a trait, will be found at a higher frequency in cases than controls. Statistical analysis is carried out to indicate how likely a variant is to be associated with a trait. The p-value [78] of the association indicates how likely the variant is to be associated with the trait (Figure 14).
Figure 14 GWAS experiments are designed to compare the frequency of variants in cases (individuals with the phenotype) than controls (individuals without the phenotype).

**Studies on the functional consequences of variants**

The majority of genetic variants that we observe are tolerated by an organism but some variants can change the function of a gene’s product. Functional genetic variation studies aim to understand the molecular mechanisms and pathways that link genotype to phenotype.

In human genetics this is an important step in translating genotype data for use in the clinic [79].

Simple variants that alter the translated protein sequence, such as, missense, splice site variant, stop gained, stop lost variants, can cause functional consequences by:

1. Altering ligand [80] and/or co-factor binding sites.

2. Alter the natural protein structure by

   a. Removing or adding additional cysteine reduces that can alter disulfide bond patterns.

   b. Alter normal formation of secondary structure [81] elements or their interaction (sickle cell anaemia is an example of this).
c. Disrupt the normal interactions between proteins' tertiary protein complexes or other cellular components.

3. Remove or add post-translational modification sites.

The functional consequences can be interpreted using prediction tools such as SIFT [82], PolyPhen [24] or Ensembl's Variant Effect Predictor (VEP) [49]. Known or predicted functional consequences for variants of a specific protein are summarised in UniProt [56].

Functional consequences due to structural variants are usually defined by the physiological phenotypes observed. These can be complex descriptions which are described using general phenotypic traits rather than specific biochemical effects caused by the variant. Clinical functional consequences are represented by a simple controlled vocabulary [83] that defines the relative pathogenicity of a variant (Table 4). Ensembl use a standard set of clinical significance terms are defined by the ACMG/AMP [84] [85].

Table 4 Controlled vocabulary terms used to describe clinical functional consequences of a variant.

<table>
<thead>
<tr>
<th>Clinical significance value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
</tr>
<tr>
<td>Likely benign</td>
</tr>
<tr>
<td>Uncertain significance</td>
</tr>
<tr>
<td>Likely pathogenic</td>
</tr>
<tr>
<td>Pathogenic</td>
</tr>
</tbody>
</table>

We will learn more about exploring the functional consequences for variants in specific proteins in the case studies [65].

Population genetics

Population genetics is the study of variation within populations of individuals, and the forces which shape it. This involves studying changes in the frequencies of genetic variation in populations over space and time.

Some of the major forces that shape variation in natural populations are: mutation, selection, migration and random genetic drift [86]. When a new mutation arises, it may be beneficial to the organism; deleterious (harmful) to the organism; or it may be neutral (have no effect on the fitness of the organism). Generally, beneficial and deleterious mutations are subject to natural selection, typically leading to increases and decreases in their allele [87] frequency, respectively.

Allele frequencies are also influenced by random genetic drift. This process explains the chance fluctuation in allele frequencies from one generation to another due to independent assortment in meiosis. This is the major force acting on neutral variants.

Modern population genetics makes use of increasing amounts of genome-scale data. It is a very large field with multiple applications. For example, data from genome-scale population genetics studies has been used to:

- Map human migration [88] [89].
- Track the spread and evolution of antimicrobial resistance [90] [91].

If you would like to learn more about the field of population genetics have a look at this section [92] of the University of Leicester's Virtual Genetics Education Centre [93].
You can browse population genetics data in Ensembl [62]. You can learn more about this in the Population Genetics section our course Ensembl: Browsing chordate genomes [94].

Summary

What is genetic variation?

Genetic variation is fundamental to the evolution of all species and is what makes us individuals. Mutations are the original source of genetic variation. In humans, recombination contributes to genetic variation by shuffling parental DNA and creating new combinations of variants.

Genetic variation is commonly divided into three main forms:

- Single base-pair substitution, also known as single nucleotide polymorphism (SNP)
- Insertion or deletion, also known as ‘indel’
- Structural variation

The term variant is used to refer to a specific region of the genome which differs between two genomes. Different versions of the same variant are called alleles. The term reference allele refers to the base that is found in the reference genome. The alternative allele refers to any base, other than the reference, that is found at that locus.

Genetic differences or variation between individuals leads to differences in an individual’s phenotype, trait or risk of developing a disease. A mendelian trait is one that is controlled by a single locus, for example a single SNP. For complex phenotypes there may be multiple variants in the genome that increase or decrease the likelihood of an individual having a certain trait, along with environmental factors.

What are the effects of genetic variation?

Variants can be categorised based on where they fall with respect to genes and other genomic features. Variant falling within a coding region are classified as synonymous, missense or nonsense variants based on how they affect the codon they falls within.

Variants in transcription factor binding sites are called TF binding site variants.

The effects of variants on protein structure can vary dramatically depending on the type of protein and the extent of variation.

Studying genetic variation

Some common steps in genetic variation studies include variant calling, variant analysis and prediction of variant effects on protein structure and function.

Variant calling is the process by which we identify variants from sequence data. It involves the alignment of whole genome or whole exome sequencing data to a reference genome in order to identify where the aligned reads differ from the reference genome. The results are stored in a VCF
file.

VCF is the standard file format for storing variation data. VCF files are tab delimited text files. Each column in the file provides information about a particular variant. VCFs are the preferred format because they are unambiguous, scalable and flexible.

**Variant identifiers** are unique combinations of letters and numbers that are assigned to known variants. Different types of identifiers are used for short variants and structural variants.

Predictions can be useful for understanding why a particular variant may cause a particular phenotype. However, prediction can never be as valuable as experimental data. Any predictions determined bioinformatically should be followed up experimentally.

Three common genetic variation study types are GWAS, studies on the functional consequences of variants and population genetics.

**What next?**

Why not take our quiz on the next page and test your knowledge from this course? After that, we recommend taking a look at part II of our human genetic variation course [9] in which we will learn how to explore publicly available genetic variation data.

**Your feedback**

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and enhance your learning experience.

**Quiz: Human genetic variation (I): an introduction**

- **Questions:** 5
- **Attempts allowed:** Unlimited
- **Available:** Always
- **Pass rate:** 75%
- **Backwards navigation:** Allowed

**References**

March Pages 67–77.
16. Population genetics in Virtual Genetics Education Centre. The University of Leicester. Last accessed 7th July 2017. URL: http://www2.le.ac.uk/projects/vgec/schoolscolleges/topics/population-genetics [92]

Contributors

David Armstrong [1]
EMBL-EBI
Scientific Database Curator - Kleywegt team: Protein Data Bank in Europe Scientific Database Curator - Velankar team: PDBe content and integration
Melissa Burke [2]

EMBL-EBI
Scientific Training Officer (e-learning): Training Team

Melissa is the Scientific Training Officer (e-learning) for the Training Team at the EMBL-EBI. She joined the Training Team in July 2016 after having worked as a Scientific Curator for ArrayExpress/Expression Atlas at the EMBL-EBI. She has a PhD in Molecular Parasitology and has worked internationally as a postdoctoral researcher specialising in the functional genomics of infectious diseases.

Laura Emery [3]

EMBL-EBI
Scientific Training Officer - Training

Laura completed her undergraduate degree in Genetics (BSc first class hons) at the University of Nottingham. After dabbling in experimental work on a summer placement at the Wellcome Trust Sanger Institute, she found joy in using computer-based analyses to explore biological questions. Laura completed her PhD in Evolutionary Genetics at the University of Edinburgh’s Institute of Evolutionary Biology. There she used a bioinformatics approach to investigate patterns of codon usage in Archaea while under the supervision of Prof Paul Sharp FRS.

She also worked as a postdoc at the University of Manchester, where she explored molecular co-evolution among interacting proteins. Having obtained qualified teacher status (QTS), Laura joined EMBL-EBI as a Scientific Training Officer in 2012 and is responsible for the development and delivery of training courses.

Jackie MacArthur [4]

EMBL-EBI
Project Leader (GWAS Catalog) - Cunningham team: Variation Annotation
Andrew Nightingale [5]
EMBL-EBI
Bioinformatician - Martin team: Protein Function development

Andrew has been a Bioinformatician with the Protein Function Development team at EMBL-EBI since 2012, and has a background in biochemistry and bioinformatics. Before joining EMBL-EBI, Andrew worked for a biotechnology company specialising in structure-based drug design by NMR. Andrew is responsible for enhancing the links between UniProt [74] and the genomics and translational biology communities to provide clinical and/or translational context to UniProtKB [103].

Emily Perry [6]
EMBL-EBI
Ensembl Outreach Project Leader

Emily is the Outreach Project Leader for Ensembl: she is responsible for the team that teaches workshops, creates training materials and help pages, manages social media, answers helpdesk queries and aids development of new tools for the resource. Emily started at EMBL-EBI as an Ensembl Outreach Officer in September 2012 and became the Project Leader in March 2015. Before working at EMBL-EBI, Emily did her PhD in molecular biology at the MRC Human Genetics Unit in Edinburgh, then worked for the University of Edinburgh's SCI-FUN group, touring Scottish secondary schools with an interactive science roadshow.
**Sangya Pundir** [7]

EMBL-EBI  
User Experience Manager - Martin team: UniProt development

Sangya Pundir is a User Experience (UX) Manager in EMBL-EBI’s UniProt team, where she established a user-centred process for the redesign of the world’s leading protein resource. To make UniProt easy for researchers to explore, Sangya conducts usability testing and information-gathering methods such as card sorting, contextual studies and workshops. Before she came to EMBL-EBI, Sangya worked at a healthcare consultancy, designing bespoke management systems. She holds an MSc in Biotechnology, Bioprocessing and Business Management from the University of Warwick.

**Gary Saunders** [8]

EMBL-EBI  
Genetic Variation Scientific Curator - Paschall team: Variation

Gary Saunders is an EBI curator of the European Variation Archive and related resources: the Database of Genomics Variants archive and the European Genome-phenome Archive. It is Gary's responsibility to manage the data within these resource(s) to ensure accuracy, clarity and discoverability. Previous to this position, Gary was a curator of the GENCODE project, which provides the gene set for the Ensembl genome browser.

Gary moved into curation following the completion of his PhD at the University of Glasgow, where he employed a variety of phylogenomic and bioinformatic methods to investigate drug resistance in nematode parasites of human and livestock importance. ORCID iD: 0000-0002-7468-0008

**Source URL:** https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction

**Links**

[1] https://www.ebi.ac.uk/training/online/trainers/davida_1  
[2] https://www.ebi.ac.uk/training/online/trainers/mburke  
[3] https://www.ebi.ac.uk/training/online/trainers/laura.emery  
[4] https://www.ebi.ac.uk/training/online/trainers/jalm_12464
Human genetic variation (I): an introduction
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[50] http://www.ebi.ac.uk/eva/
[53] http://cancer.sanger.ac.uk/cosmic
[57] http://www.ebi.ac.uk/dgva
[60] https://vcftools.github.io/
[64] https://www.ebi.ac.uk/gwas/
[65] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-exploring-publicly-available-data/case-studies
[66] https://xkcd.com/1430/
[67] http://bleoberis.bioc.cam.ac.uk/duet
[70] http://zhanglab.ccmb.med.umich.edu/l-TASSER
[71] http://www.sbg.bio.ic.ac.uk/~phyre2
[72] http://PDBe.org
[73] https://www.ebi.ac.uk/training/online/glossary/pymol
[74] https://www.ebi.ac.uk/training/online/glossary/uniprot
[75] http://www.ebi.ac.uk/training/events/2016/pdbe-searching-biological-macromolecular-structures
[76] https://www.ebi.ac.uk/training/online/glossary/single-nucleotide-polymorphism
[77] https://www.ebi.ac.uk/training/online/glossary/ secondary-structure
[78] https://www.ebi.ac.uk/training/online/glossary/p-value
[79] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/##genvar_ref_12
[80] https://www.ebi.ac.uk/training/online/glossary/ligand
[81] https://www.ebi.ac.uk/training/online/glossary/secondary-structure
[83] https://www.ebi.ac.uk/training/online/glossary/controlled-vocabulary
[85] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/##genvar_ref_13
[86] https://www.ebi.ac.uk/training/online/glossary/random-genetic-drift
[87] https://www.ebi.ac.uk/training/online/glossary/allele
[89] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/##genvar_ref_14
[90] http://europepmc.org/articles/PMC3705945/
[91] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/##genvar_ref_15
[92] http://www2.le.ac.uk/projects/vgec/schoolscolleges/topics/population-genetics
[93] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/##genvar_ref_16
[94] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/navigating-ensembl/investigating-sequence-variation/varia-2
[95] http://www.nature.com/scitable/definition/allele-48
[99] http://europepmc.org/abstract/MED/28213901
[100] http://europepmc.org/abstract/MED/25741868
[103] https://www.ebi.ac.uk/training/online/glossary/uniprotkb