Background scientific knowledge

Table of contents

Background to the workshop

Introduction

The purpose of this document is to provide staff responsible for delivering *Hands-on DNA: Bacterial Evolution* workshops with the necessary background information required to deliver the workshop successfully. It has been written for a general scientific audience who may have little previous experience of genetics or evolution.

Project background: *A Question of Taste*

A Question of Taste is a workshop developed as part of a programme of educational projects celebrating the bicentenary of Darwin's birth, and 150 years since the publication of his most famous work, *On the Origin of Species by Means of Natural Selection*. Nowgen, At-Bristol and Centre for Life originally developed the workshop in the UK, on behalf of the Wellcome Trust.

Hands-on DNA: Exploring Evolution

Hands-on DNA: Exploring Evolution is a project led by The Association for Science and Discovery Centres (ASDC), in collaboration with the original project partners (Nowgen, At-Bristol and Centre for Life), again supported by the Wellcome Trust. The project's vision is to give students across the UK the opportunity to explore evolution by providing access to high-quality engaging molecular biology experiences. *Hands-on DNA: Exploring Evolution* will make this possible by providing training and equipment to enable new organisations to deliver either the *A Question of Taste* PCR workshop or *Bacterial Evolution,* a new molecular biology workshop.

Overview

Hands-on DNA: Bacterial Evolution is based on the genetic techniques scientists use to identify bacteria and illustrates how, by looking at the DNA of bacteria, we can see evolution happening in real time.

In particular the workshop investigates the *emm* gene of *Streptococcus pyogenes* which encodes an important virulence factor, the M protein of this human pathogen. The *emm* gene is highly variable and this results in over 250 variants of the M protein, called M types. Diseases caused by *S. pyogenes* range from impetigo and strep throat to Streptococcal toxic shock syndrome (STSS) and necrotizing fasciitis. While the former are relatively common and non-severe, they do nonetheless place a burden on health services and costs billions of dollars a year to treat. The latter are serious diseases, which are of great risk to infected individuals.

The M type is one of the key determinants as to whether an individual infected with *S. pyogenes* will develop strep throat or, for example, STSS. It is clearly important to be able to identify the M type so that appropriate prophylaxis or treatment can be given. The M type can be quickly determined using DNA technology to sequence the hypervariable portion of the *emm* gene most responsible for giving rise to the different M protein variants. In outbreak situations where

access to DNA sequencing technology is limited, restriction fragment length polymorphism (RFLP) analysis can be used to distinguish between different M types and students carry out a simplified version of this technique in the workshop.

The M protein (and therefore *emm* gene) represents a classic evolutionary arms race. One of its key roles is evasion of the immune system. The adaptive immune system in humans can be viewed as an independently evolving system in every individual – it evolves to recognise and destroy pathogens as it encounters them throughout life. As such, pathogens have to fight back. There is selective pressure on bacteria that are easily recognised and destroyed by the immune system and they will be removed from the population. Conversely, the population of bacteria with slightly different characteristics (in this instance, with a new form of the M protein) will continue to expand.

How does this build on Darwin's work?

Darwin proposed that all species have evolved from common ancestors through a process termed natural selection. In his most famous work *On the Origin of Species by Means of Natural Selection* the only diagram is a tree of life, which shows the relationships between species that have evolved, in response to selection pressures, from common descent. This workshop considers the evolutionary pressures that act on the most abundant organisms on Earth: bacteria. In particular, it looks at how the bacterial species *Streptococcus pyogenes* has evolved variants in one of its genes in response to pressure from the human immune system.

Darwin's work has been supported by fossil evidence and comparative studies in anatomy, physiology and biochemistry. Genetic information provides further supporting evidence for his theories. As DNA sequencing techniques and computing capability continue to advance, the scientific community has access to an increasing body of DNA sequence data with which to study evolution. Projects are underway to collect genetic information from all of the world's species which can then be compared using internet-based bioinformatics software. Recently, the particularly virulent outbreak of *E. coli* that started in Germany[1] led to the novel bacterial strain having its genome sequenced within a matter of days of the outbreak occurring. This allowed bacteriologists to look at the key parts of its DNA code that made it different, and more dangerous, than other *E. coli* strains.

A workshop that bridges evolutionary research and molecular biological techniques

The practical component of the workshop allows students to examine real DNA using a restriction digest and gel electrophoresis, and consider how bacteria can evolve to evade the human immune system and cause disease. Students use research-grade equipment, including micropipettes, to improve their practical skills and the workshop is fully curriculum-linked,

[1] http://www.bbc.co.uk/news/health-13600144

making it attractive to teachers. Finally, and perhaps most importantly, the workshop demonstrates how a molecular technique is applied to a genuine medical problem – how to rapidly and accurately identify a particular strain of bacteria. Students will learn the theory of the techniques, and apply them practically to investigate DNA. This work brings together the history of evolution and cutting edge genetics research.

How does the workshop work?

Hands-on DNA: Bacterial Evolution takes advantage of pre-prepared DNA samples and an easy-to-use restriction enzyme that cuts DNA in a stereotypical and predictable way. These fragments are separated and visualised using gel electrophoresis.

Students are provided with a DNA sample from one of four outbreaks of *S. pyogenes* and two reference samples from known M types that cause mild and severe streptococcal disease respectively. They use a lyophilised (effectively, freeze-dried) restriction enzyme to cut the samples, which generates a number of fragments of different sizes. The students compare the pattern of fragments in the three samples and predict which kind of disease their outbreak sample might cause and also look out for any new M characteristics revealed by their analysis.

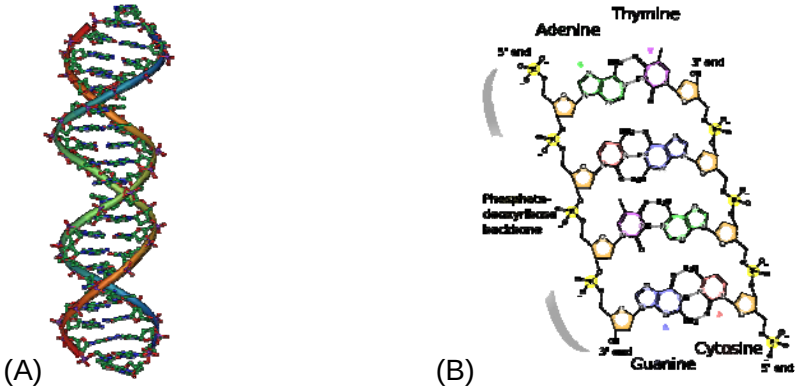
The workshop highlights that there are genes that vary between individual organisms of the same species and that selective pressure acting on these genes provides a mechanism for evolution.

A brief introduction to genetics and molecular biology

Genetics is the study of genes, variation and inheritance, while molecular biology is concerned with the molecular basis of biological processes. Both disciplines overlap considerably.

DNA

DNA is 'the molecule of inheritance' and its unique structure allows it to encode the information for other chemicals that help your body to grow, develop and function. Famously, the structure of DNA is that of a 'double helix' (Figure 1), which is akin to a twisted ladder.



(A) (B)

Figure 1 (A) DNA in a double helix. You can clearly see its backbone and the bases projecting inwards. (B) Simplified view of DNA, illustrating its constituent parts and showing the backbone, the four bases and how they link together. Both pictures are from Wikipedia.

DNA is made from two long chains of repeating subunits called nucleotides. A nucleotide has three parts:

1. a phosphate group
2. a sugar (deoxyribose)
3. a base

The sugar and phosphate groups link together to form DNA's sugar-phosphate backbone. The bases are connected to the sugar and point inwards, where they bind via hydrogen bonds with a complementary base on the second strand, forming a base pair. DNA contains four bases:

1. Adenine
2. Cytosine
3. Guanine
4. Thymine

For ease these bases are usually just referred to by their first letter; so, 'A', 'C', 'G' and 'T'. Accordingly, the bases are often called the 'letters' in DNA's 'alphabet'. A base in one strand of DNA always has a complementary base in the other strand: 'A' always complements 'T' (and vice versa), while 'C' always complements 'G' (and vice versa).

The chemistry of DNA is such that each strand can be described as having distinct ends, which are called the 5' (said, 'five prime') and 3' ('three prime') ends. The strands are 'anti-parallel', which means that one strand runs 5' to 3', while its complementary strand runs 3' to 5' (see Figure 1B). The bases that ultimately form the code of DNA can be read along a length of DNA; by convention, and to avoid confusion, DNA is always read 5' to 3'.

The difference between the DNA of one human and another is tiny; any two humans, at the DNA level, are over 99.9% identical. Many of the differences are accounted for by changes in single bases between individuals. For any given nucleotide position in the genome, most will be the same, but in some instances there will be a 'choice' of bases. For example, if we looked at nucleotide 58,358,123 (to pick a very random number!) 20% of the population might have a 'T', while 80% of the population might have a 'G'. These variations are known as polymorphisms (specifically, in this case, a single nucleotide polymorphism). An individual's (or an organism's) set of polymorphisms is called its genotype. A genotype can either refer to the polymorphisms of an entire genome, or a smaller, discrete portion of the genome all the way down to a single nucleotide.

Genes and proteins

Genes can be described as discrete portions of DNA which encode another chemical, usually a protein. Proteins are the workhorses of a cell; they 'make things happen'. Like DNA, proteins are polymers, but rather than being made of nucleotides they are made of amino acids. The sequence of amino acids in a protein is determined by the sequence of bases in a gene. In the DNA three bases form a codon, which encode a single amino acid. Note that one codon only encodes one amino acid, but some amino acids are encoded by more than one codon; this is because there are only 20 amino acids, but 64 possible codons (4 possible bases at position 1 in the codon x 4 possible bases at position two x 4 possible bases at position three = 64).

Because DNA determines the sequence of amino acids in a protein, it follows that a change in the DNA could result in a change to a protein. It is changes to proteins that usually cause an observable change to an organism; that is, a change in phenotype.

To make a protein from a gene, the gene is first copied to an intermediate molecule called messenger RNA (mRNA) in a process called transcription. Like DNA, mRNA is also made of nucleotides that contain bases, but in RNA the sugar is slightly different. Also, in RNA the base 'T' is replaced by another base, uracil ('U'). Because of the nature of DNA, the RNA copy is complementary to the template DNA strand. Following transcription, mRNA is transported from the nucleus (where the DNA resides) and translated into proteins by molecular machines called ribosomes. Ribosomes work by sequentially reading the codons in the mRNA and incorporating the appropriate amino acid for each codon. There are four special codons which act as punctuation, one is the start codon (ATG in DNA; AUG in RNA), the other three are stop

codons. The start codon tells the ribosome where to start building the amino acid chain; a stop codon tells the ribosome where to stop.

Genes often exist with multiple alternative versions, called alleles. A good example is human blood groups, controlled by the ABO gene. Three different versions (alleles) of ABO exist denoted as A, B and O. Every individual inherits one copy of the gene from each parent, so for the ABO gene, there are six possible genotypes: AA, AO, BB, BO, AB and OO (AO is equivalent to OA as is AB and BA and BO and OB, hence why there are not nine genotypes).

If you want to find out more about genetics, you might like to look at:

- The DNA Learning Center's, *DNA from the beginning* http://www.dnaftb.org
- The University of Utah's, *Genetic Science Learning Centre* http://learn.genetics.utah.edu/

A brief introduction to evolution

Evolution describes the phenomenon of organisms' inherited characteristics changing over time. Such changes are typically small and happen gradually but, in combination with natural selection and speciation, evolution led to simple, single-celled organisms giving rise to millions of different, weird and wonderful complex forms of life.

Evolution relies on changes to DNA. Alterations to DNA lead to differences in an organism and these differences being passed on from one generation to another. Natural selection acts to preferentially maintain those organisms (or individuals) that have characteristics suited to their environment, while removing organisms/ individuals with less well-suited characteristics. Beneficial changes to an organism will therefore spread throughout its population, while harmful changes will not.

Speciation occurs when two populations of an organism are isolated in some way and are therefore able to develop independent changes, which do not spread between populations. Eventually, the two populations will develop characteristics so distinct (and, by implication, have such different genetics) that they can no longer interbreed and can be considered separate species. (Note, other, hotly debated, mechanisms of speciation exist, but they are beyond the scope of discussion here.)

In order for evolution to occur it is important that DNA is not immutable. Were DNA unable to change (mutate) there would be no way that the phenotype could alter. DNA is able to change in a variety of different ways, including spontaneous, small mutations, called 'point mutations' where just a single base pair is changed. A mutation can be:

- **neutral** and have no observable effect;
- **positive** and give the individual with that mutation a reproductive advantage (and therefore more likely to pass on their genes and the mutation);
- **negative** and put the individual with that mutation at a disadvantage (and therefore less likely to pass on their genes and the mutation).

It is important to bear in mind that a mutation that is initially neutral might, at some point in the future, become either positive or negative.

Natural selection acts upon the phenotype of an organism, and increases or decreases the likelihood of an individual passing on their genetic information to the next generation, depending on the effect of the phenotype. Because the phenotype is the manifestation of the genotype, natural selection, by definition, affects the distribution of genes, and therefore alleles, in a population.

More information about evolution is available online from the University of Berkley's Understanding evolution: http://evolution.berkeley.edu/

A brief introduction to bacteria

Bacteria are among the oldest single-celled organisms to have come into existence. It is estimated they first evolved around 3.5 billion years ago. Bacteria form one of the so-called three 'domains' of life:

- Bacteria – small, simple, single-celled organisms whose DNA is free within the cell, and not contained within a nucleus.
- Archaea – small, simple, single-celled organisms, similar to bacteria but with different biochemistry.
- Eukaryota – all other forms of life; their DNA is contained within a discrete nucleus inside their cell(s). Includes some single-celled organisms such as amoebae, but also complex, multi-cellular organisms such as plants and animals.

Bacteria and Archaea are very similar; confusingly, some Archaea have Latin names that contain the word 'bacteria'. This is because until as recently as 1990 bacteria and Archaea were classified together as prokaryotes. However, recent analysis of the two showed that there were sufficient differences between them to warrant classifying them separately. The eukaryota domain evolved out of the Archaea; scientists' current thinking of the evolution of life is illustrated in Figure 2.

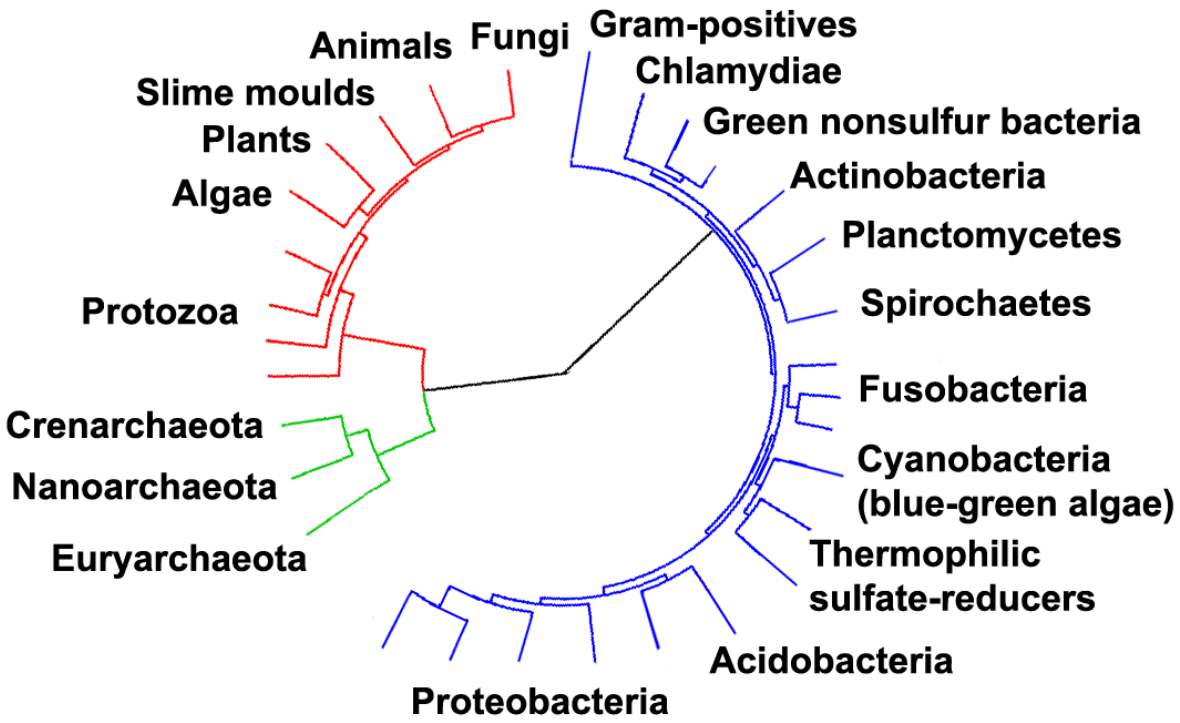


Figure 2 Illustration of how life evolved. The central point of the circle represents the 'last universal ancestor'. Bacteria are shown in blue; eukaryotes are shown in red; Archaea are green. Only broad divisions are highlighted. *Streptococcus pyogenes* (the subject of the workshop) are gram-positive bacteria.

Bacterial cells are considerably smaller and simpler than eukaryotic cells. Typically, they range between 0.5-5µm – about one tenth of the size of a eukaryotic cell. Bacteria have a range of different shapes, but most are either spherical (the 'cocci' such as *Streptococcus*) or rod-shaped (the bacilli, such as *E. coli*).

The genome (total genetic information) of bacteria is very simple (at least in comparison to eukaryotic genomes), consisting of just a single chromosome in the form of closed loop of DNA, ranging in size from just 580,000 base pairs (see *A brief introduction to genetics and molecular biology*, above) to approximately ten million base pairs. For comparison, the human genome has three billion base pairs divided into 23 chromosomes of linear DNA: the bacterial chromosome is very different in structure to the chromosomes found in humans.

Because bacteria contain just a single chromosome, they are described as haploid; unlike humans who have two copies of each of their 23 chromosomes and are therefore diploid. That bacteria are haploid is important, because it means that any change in a coding region of their DNA will have an effect. In diploid organisms a change in the DNA of one chromosome might not necessarily lead to an effect, since the change might be masked by the corresponding DNA of the sister chromosome.

In addition to their single chromosome, bacteria also frequently contain even smaller closed loops of DNA of just a few thousand base pairs. These smaller loops are called plasmids. Plasmids are important because they allow for horizontal gene transfer. This is where genetic information is passed from one bacterial cell to another, without one cell being a descendant of the other. (Vertical gene transfer refers to genetic information passed from parent to offspring.) This allows plasmid DNA to be shared rapidly through an entire population of bacteria and is particularly important as mechanism of spread of antibiotic resistance. Plasmids are also of note because they are a staple tool of molecular biologists; they can easily be manipulated and engineered in a laboratory, allowing genes to be examined and foreign genes to be inserted into bacteria.

Bacteria are perhaps the most adaptable and populous of all species. There are very few environments on earth where they are not found, ranging from freezing ice to hydrothermal vents, from your gut to your skin. In terms of numbers it is estimated that:

- 1 gram of soil contains approximately 40 million bacterial cells;
- 1 millilitre of water contains approximately one million bacterial cells;
- in total there are $5x10^{30}$ bacterial cells on Earth – their collective biomass is greater than that of all plants and animals combined; and
- there are around ten times more bacterial cells in and on your body than cells of your body.

Of the bacteria that live within or on humans they can be innately harmless, rendered harmless by the immune system, and some are actively beneficial. However, occasionally, bacteria are able to overcome the immune system and cause disease. Such bacteria are both parasitic and pathogenic.

Streptococcus pyogenes

One species of bacteria that is sometimes found living on or in humans is *Streptococcus pyogenes*, abbreviated to *S. pyogenes*, and also called Group A streptococcus (GAS). It is estimated that 5-15% of people harbour *S. pyogenes*, usually in their respiratory tract (the lining of the lungs etc.). *S. pyogenes* does not usually cause disease, but can when an individual's immune system is compromised, typically due to infection by another pathogen, such as a virus. Typically, diseases caused by *S. pyogenes* are relatively mild, the most well-known being 'strep throat' and impetigo (although impetigo is more commonly caused by another bacterium). Non-specific defences in humans, such as the skin, coughing and sneezing are all effective at preventing the bacteria passing too far into the respiratory tract and causing disease.

Should *S. pyogenes* manage to get within the body, and therefore the bloodstream it can cause more serious, 'invasive disease'. Notably, it causes the now infamous necrotizing fasciitis – often dubbed 'flesh eating bacteria' by the media. It addition it can cause streptococcal toxic shock syndrome (STSS), which is essentially the body's immune system going into overdrive and producing large amounts of toxic chemicals. Usually, these chemicals are produced in small amounts to help kill off an infection, but in STSS the large amounts produced can lead to fatal toxicity within the body. Approximately 20% of people with necrotizing fasciitis, and over 50% of people with STSS die as a direct result of infection with *S. pyogenes*.

Bacterial cells within the body can be identified by two key parts of the immune system: the complement system and antibodies – see *A brief introduction to the human immune system*, below. Both effectively mark bacterial cells for destruction by phagocytes ('eating cells' – cells of the immune system that engulf and destroy bacterial cells). *S. pyogenes* infections are also effectively treated with penicillin.

S. pyogenes remains a clinically significant pathogen due to the combination of the large number of mild infections it causes, and the small number of very serious infections. In the US there are between 1,000 and 1,800 deaths annually as a result of invasive disease caused by *S. pyogenes*. In addition, the millions of incidences of strep throat and impetigo cost billions of dollars in medical expenses.

The 'M' protein and *emm* gene

A key gene in the pathogenicity of *S. pyogenes* is the *emm* gene, which encodes the 'M' protein. The M protein, the product of the *emm* gene, protrudes from the surface of the bacterial cell and

helps the bacterium to attach itself to host cells. There are over 250 different versions of the M protein, with the different variants being known as M types. The different M types confer different properties on the bacteria, in particular its virulence (that is, how effective it is at causing disease). Consequently, knowing the M protein type is important when dealing with *S. pyogenes* infection. Direct analysis of the bacteria's DNA allows the *emm* gene to be examined and therefore the M type deduced. This is known as *emm* typing and it particularly examines one region of the DNA, responsible for encoding the part of the protein that is most variable (known as 'hypervariable').

The M protein is important because the different versions can confer certain advantageous characteristics to the bacteria. The most notable advantage is that of being able to avoid phagocytosis; that is, it prevents phagocytes from engulfing and destroying bacterial cells. This occurs because some variants of the M protein are able to prevent proteins in the complement cascade marking bacterial cells for destruction. Interestingly, different variants of M proteins inhibit this process in different ways, highlighting the importance of M proteins in fulfilling this function. The M protein can also help bacterial cells being destroyed by phagocytosis by binding with antibodies in such a way that the phagocytes are unable to recognise them and respond accordingly (that is, destroy the bacterial cell).

Antibiotic resistance

When Alexander Fleming discovered the world's first antibiotic, penicillin, in 1928 many hailed it as an end to infectious disease. Of course, this hasn't come to pass (for a start, antibiotics only work on bacteria, not viruses), and instead many bacteria have started to evolve resistance to antibiotics.

Anything that rapidly kills an organism puts a huge selective pressure on that organism to respond and adapt. Bacteria reproduce incredibly quickly[2], which allows them to adapt quickly to such selective pressures. This is the case with antibiotic resistance – just one individual bacterium needs to evolve to be resistant, and that resistance will quickly spread throughout the entire population (see FIG).

The flip side is that drug resistance is 'difficult' to evolve, since drugs often target crucial biochemical pathways that have been finely tuned over millennia and are therefore intrinsically resistant to change. However, resistance only needs to evolve once to spread and have far-reaching consequences. In fact, genes for antibiotic resistance pre-date the use of antibiotics in medicine, but the use of medical antibiotics has exacerbated that selective pressure and leading to an increase in antibiotic resistant strains.

[2] Under optimum conditions, some bacteria can divide and reproduce every 20 minutes; in theory, in one day, a single bacterium could reproduce to form a colony of over $4x10^{21}$ individuals – vastly more than the number of humans that have ever lived. In reality does not happen since the growth rate slows to zero as the population density increases.

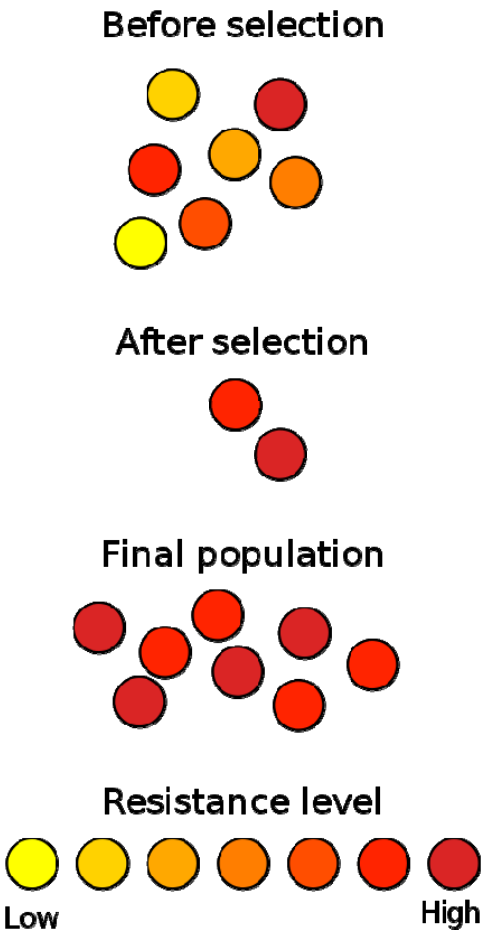


Figure 3 Development of antibiotic resistance in bacteria. In this instance, 'selection' is equivalent to giving an antibiotic – it kills bacteria with low resistance, effectively selecting those with high resistance. Image from Wikipedia: http://en.wikipedia.org/wiki/Antibiotic_resistance

Often, when bacteria do become resistant to one type of antibiotic, they remain sensitive to other antibiotics. However, it is possible for bacteria to be multi-drug resistant, particularly when resistance is conferred by genes residing in plasmids that may be spread by horizontal transfer. Multi-drug resistant bacteria are often referred to colloquially as 'superbugs'. The antibiotic, 'vancomycin' remains one of the most powerful and effective antibiotics, but some bacteria are even resistant to that.

As noted above, antibiotics are only effective on bacteria, not viruses. Nonetheless, antibiotics are frequently prescribed for non-bacterial infections, or for relatively mild infections that the immune system should be able to deal with. It is the amount of antibiotics that is prescribed that is a key factor in the development of antibiotic resistance (more so than non-compliance with a course of treatment).

A brief introduction to the human immune system

Without an immune system your body would quickly be overcome by infection and you would almost certainly die with a matter of days. Not only does this not happen, you rarely even register the fact that your immune system is constantly, quietly removing invaders, keeping you healthy. Only when a particularly virulent pathogen comes along does your immune system have to markedly increase its activity, resulting in the production of chemicals that help kill off the infection, but also make you feel unwell.

The key aspect of your immune system is its ability to distinguish between what is you ('self') and what is not you ('non-self'). Broadly, it achieves this through two 'arms':

- the innate, or non-specific, immune system;
- the adaptive immune system.

As their names suggest, the two arms work in different ways. The innate immune system does not respond to specific pathogens, but recognises and responds to the generic characteristics of a range of pathogens. Conversely, the adaptive immune system only acts on pathogens once it is able to selectively identify them. As such, your adaptive immune system has to 'learn' before it is effective, whereas your innate immune system works 'straight out of the box'.

Your innate immune system cannot recognise the wide range of pathogens that your adaptive immune system can – it is estimated that your adaptive immune system can recognise up to a billion different antigens (see below). After the adaptive immune system encounters a pathogen it 'remembers' and, should the body be infected again by that same pathogen, responds more rapidly and to a greater extent second time around (Figure 4).

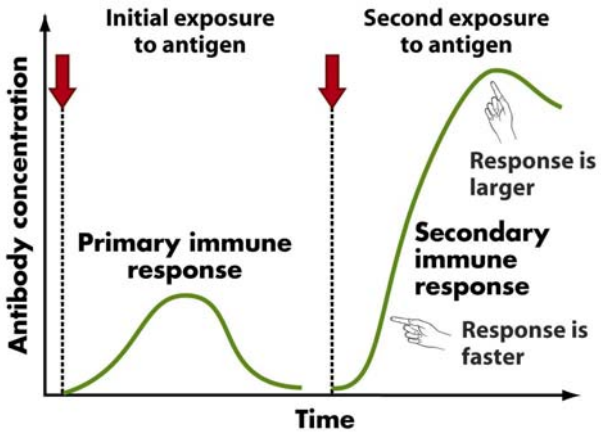


Figure 4 The adaptive immune response[3] means that your body 'remembers' previously encountered pathogens and deals with them quicker and more efficiently if it encounters them again. 'Antibody concentration' (on the *y* axis) is equivalent to 'size of immune response'.

[3] Image taken from http://www.uic.edu/classes/bios/bios100/lectures/memory.jpg

One of the important components of the innate immune system is called the ‘complement system’. This comprises a number of proteins which can interact, non-specifically, with a range of pathogens. When complement proteins bind to pathogens they become activated and alert a special type of cell of the immune system called phagocytes. Phagocytes are large cells that engulf smaller pathogens, in particular bacteria, as shown in Figure 5. (The name phagocyte literally means ‘devouring cell’.) In general, your innate immune system refers to anything that acts to prevent infection by pathogens; so this could include your skin or the mucus that lines your lungs.

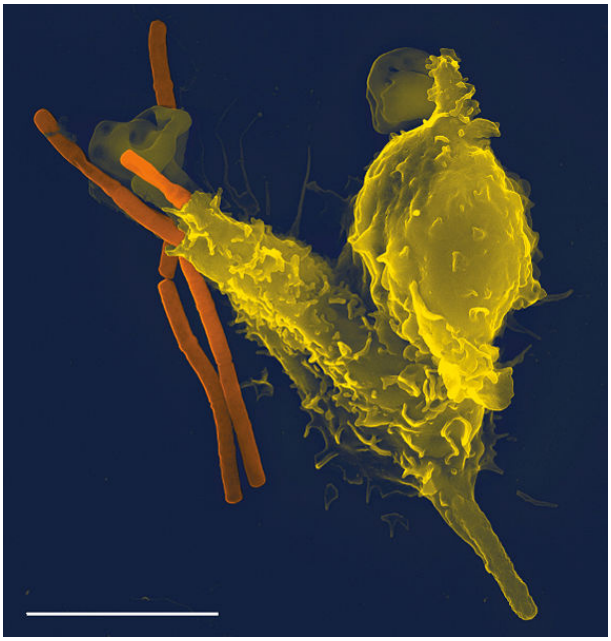


Figure 5 Phagocytosis of anthrax (a bacterium) by a phagocyte. The phagocytes are the large, rough cells coloured in yellow; anthrax the smooth rod coloured orange. The white line is a scale bar showing 5µm (five one thousandths of a millimetre) – the anthrax bacteria are very small!

Your adaptive immune system is based upon recognising unique signatures called ‘antigens’ that are present on pathogens. By definition, antigens must activate the immune system and cause an immune response. Specialised cells of the adaptive immune system are able to recognise a huge number of different, specific antigens. Antigens are recognised by antibodies, which attach to an antigen and then recruit phagocytes to remove the pathogen containing the antigen; similarly to the complement system but in an antigen specific manner.

There is a constant arms race between pathogens and your immune system. In order to live and reproduce, pathogens must effectively infect a host organism and evade their immune system. As a result, pathogens often have key genes (especially genes whose products can be antigens) that are extremely variable (so-called hypervariable genes). However, as mentioned above, the adaptive immune system is hugely adaptable, with its own set of hypervariable genes that can quickly respond to the changes in a pathogen’s repertoire of antigens.

Glossary

Adaptive immune system	The arm of the immune system that responds selectively to certain pathogens by recognising specific antigens.
Allele	Alternative form of a *gene*. The difference between alleles can vary, from a single *SNP* to larger differences.
Amino acid	The building blocks of *proteins*; humans have 22 amino acids. The sequence of amino acids in a protein is determined by information encoded within *genes* in *DNA*.
Antibody	A special protein produced by certain cells in the adaptive immune system. Antibodies recognise and attach to specific antigens on pathogens, before recruiting other cells of the immune system to kill the pathogen.
Antigen	A substance or molecule that causes an immune response by stimulating the production of specific antibodies.
Arms race	Often used to describe the *evolution* of advantageous characteristics by a *pathogen*, which help it to overcome a host's immune system, and vice versa.
Bacterium/ bacteria	Small, simple, single-celled organisms whose *genome* is contained in a relatively small, closed loop of *DNA*. Bacteria are among the oldest and most adaptable forms of life found on Earth.
Base and base pair	A 'letter' of *DNA* and part of a *nucleotide*. DNA has four bases, adenine (A); thymine (T); cytosine (C) and guanine (G). DNA is double-stranded, so every base has a 'complementary' counterpart and is part of a 'base-pair'. 'A' always complements 'T' and 'C' always complements 'G'. The words 'base' and 'nucleotide' are often used interchangeably.
Bioinformatics	The use of computers and mathematics to analyse biological data. Most frequently used these days in the context of using computers to analyse DNA sequence data.
Cell	The basic unit of almost all life: a fluid-filled sac that is surrounded by a thin, fatty membrane. Some organisms, such as bacteria or yeast, are just a single cell; others are much more complex and made of trillions of cells, such as humans. With very few exceptions, cells contain a *genome* and in 'higher organisms' the genome is organised into *chromosomes*, which are found within the *nucleus* of the cell.

Chromosome	A threadlike structure that contains a discrete portion of *DNA*. Humans have 23 pairs of chromosomes: 22 pairs of non-sex chromosomes ('autosomes') and two sex chromosomes (XX or XY), giving 46 chromosomes in total. One chromosome of each pair comes from an individual's mother; the other comes from their father. Bacteria have a single chromosome, inherited directly via asexual reproduction.
Coding DNA	DNA within a genome which encodes a functional chemical, usually a protein.
Codon	A sequence of three *bases*, which encode a single *amino acid*. Given that a codon is three bases long and there are four possible bases at each position, there are $4^3 = 64$ different codons. See also *start codon* and *stop codon*.
Complementary	Describes the fact that every *base* on one strand of *DNA* has a complementary base on the other strand. 'A' complements 'T' (and vice versa) and 'C' complements 'G' (and vice versa).
Diploid	Refers to cells or organisms that have two sets of *chromosomes*, such as humans, who have one maternal set and one paternal set.
DNA	Deoxyribonucleic acid. A chemical made out of repeating pairs of *nucleotides*, which join up to create a very long, very thin structure that adopts the form of a *double-helix*. DNA is the chemical that carries the genetic information in almost all living things.
Double helix	The structure adopted by a length of double-stranded *DNA*. A double helix is the shape you would get by twisting the ends of a ladder.
Electrophoresis	A technique that separates fragments of *DNA* according to their size. Typically, an electric current is passed through a buffer in which sits an agarose gel containing the DNA to be separated. Large fragments of DNA migrate slowly through the gel; small fragments migrate quickly.
Enzyme	A special type of *protein* that speeds up chemical reactions. Enzymes have very specific functions.
Evolution	The process whereby, over time, new species form through the action of *natural selection* on random *mutations* which lead to favourable adaptations.

Expressed/ expression	Refers to whether a *gene* is 'turned on' and being *transcribed*. This is a very common word in genetics/ molecular biology. Some examples would include: 'Expression of the emm gene leads to the M protein being present on the surface of S. pyogenes,' or 'Expression of any gene is carefully regulated by cells'.
Gene	A portion of *DNA* that encodes another chemical, usually a *protein*. The human genome encodes around 23,000 genes. Genes are found at discrete *loci* along the length of a *chromosome*.
Genetic code	The set of rules whereby genetic information is *translated* into a sequence of *amino acids*, which form a *protein*. The genetic code defines how each *codon* is translated.
Genome	An organism's complete set of genetic information. A bacterium's genome is relatively simple, consisting of a single close loop of DNA.
Genotype	A description of a portion (be it a small part, or all) of an individual's *DNA*.
Haploid	Refers to cells or organisms that have a single set of *chromosomes*. Bacteria and viruses are haploid, and so are the sex cells (sperm and eggs) of humans.
Immune response	Increased activity of the immune system, which involves increased proliferation of cells of the immune system as well as production of special chemicals that help remove the pathogen.
Immune system	An important system in complex organisms, especially vertebrates, which helps destroy infectious pathogens such as bacteria and viruses. The key function of the immune system is to distinguish between 'self' and 'non-self' and destroy anything recognised as non-self.
Innate immune system	Refers to the non-specific, non-adaptive arm of the immune system that recognises generic features of pathogens and acts to neutralise them. More generally, the innate immune system includes any feature that helps prevent infection, such as the skin or the mucosal lining of the lungs.
Locus/ loci	A locus is part of a *chromosome* that contains a *gene* (plural: loci).
Messenger RNA	mRNA; an intermediate chemical, which is *transcribed* from *DNA* before being *translated* into a *protein*.

Mutation	A change in *DNA*. There are many different types of mutation, some which affect single nucleotides, others which affect large stretches of DNA.
Natural selection	The preferential selection of some individuals of a species over others due to slightly different characteristics which are favourable in the selected individuals. Natural selection ensures that individuals that are best adapted to their environment survive and breed, passing on their genes to the next generation.
Non-coding DNA	*DNA* within the *genome* that is not known to encode any other chemical.
Nucleotide	A repeating subunit that makes up *DNA*. Nucleotides include a phosphate group, a sugar (deoxyribose) and a '*base*'. The words 'nucleotide' and 'base' are often used interchangeably.
Nucleus	A discrete structure within a eukaryotic *cell*, which contains the cell's *chromosomes*.
Pathogen	Any virus or bacterium that is capable of causing disease.
Phagocyte	A specialised cell of the immune system that engulfs (or eats!) invading pathogens, particularly bacteria, and then destroys them.
Phagocytosis	The process of a phagocyte engulfing a pathogen.
Phenotype	The physical manifestation of a *genotype*.
Polymorphism	Literally, 'many forms'. In the context of *DNA*, often used to refer to a *nucleotide* position that has one or more different versions in different individuals.
Protein	A complex chemical made up of a string of *amino acids*, folded into a precise three-dimensional shape. Proteins are encoded by *genes* in *DNA*.
Restriction enzyme	A specialised *enzyme* that cuts *DNA* according to a specific sequence. For example, HaeIII will only cut DNA at the sequence CCGG. Restriction enzymes may leave either blunt ends or *sticky ends*, depending on how they cleave the DNA.

Restriction fragment length polymorphism (RFLP) analysis	A technique whereby *DNA* samples are cut with one or more *restriction enzymes* to generate distinctive patterns of different-sized DNA fragments (so, *fragment length polymorphisms*). Samples can be compared with each other or to reference samples, and those with identical or similar patterns of fragments will have identical or similar DNA sequences.
RNA	A chemical similar to *DNA*, but with some important differences. RNA is (usually) single-stranded, and instead of a 'T' (thymine), RNA uses 'U' (uracil). There are different forms of RNA, including messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).
Selective advantage	An adaptation or *mutation* that means that the affected individual is more likely to breed and pass it on to the next generation.
Speciation	The process of a new species forming. This happens when an ancestral population is separated in some way into two distinct populations which can evolve independently of each other. Eventually, they become so distinct, morphologically and/ or genetically that they are separate species.
Start codon	A special *codon* sequence that indicates the beginning of a protein-encoding gene. In most organisms there is only one start codon, which is ATG.
Sticky end	A feature of *DNA* that has been cut by a certain type of *restriction enzyme*. One strand of DNA slightly extends (by just a few *nucleotides*, usually 3) beyond the other strand, forming an overhang.
Stop codon	A special *codon* sequence that indicates the end of a protein-encoding gene. There are three stop codons in humans.
Transcription	The process of copying a strand of *DNA* into a molecule of *mRNA*.
Translation	The process of reading the *codons* in an *mRNA* molecule and 'translating' them into *amino acids*, thus creating a string of amino acids which ultimately form a *protein*.
Virulent/ virulence	Refers to a *pathogen's* ability to causes disease. Highly virulent pathogens cause more severe disease.
Virulence factor	A chemical made by a *pathogen* that makes them virulent. The M protein of *S. pyogenes* is a virulence factor.

Other online glossaries are widely available and may be of use. You might like to look at:

- http://insidedna.org/content/?tag=glossary
- http://www.yourgenome.org/glossary/
- http://www.geneticalliance.org.uk/glossary.htm
- http://www.phgfoundation.org/pages/resources/glossary.htm