WESTERN SYDNEY UNIVERSITY

300820

GENES, GENOMICS AND HUMAN HEALTH

School of Science

2021 Online practical classes

Complex and Mendelian Genetics: An Introduction

Introduction and Overview

What do we understand about the genetic basis of human health and disease? In this unit we will move away from Mendelian ideas of human phenotypes and start to consider more complex ideas about the genetic basis of human phenotypic variation. Essentially, this involves looking at the genetic basis of disease in populations, rather than at the level of individuals and families. When we consider phenotypes at a population level, we are looking at phenotypes that are clearly apparent in that population. To be more specific, we are looking at the variation in phenotypes that are common in a population. Examples of such phenotypes are height, body mass index (obesity) and susceptibility to allergy. Each of these phenotypes are affected, to varying degrees, by environmental and genetic factors. Changes in diet, lifestyle and hygiene are all examples of environmental changes in Western societies over the past century.

The role that environment plays in phenotypic variability is typical of complex genetic traits and is in contrast to a classical Mendelian trait. In its simplest form, a Mendelian trait is a genetic change in one gene that results in a change in phenotype. The genetic change in a gene is called an allele – humans have a diploid genome with two copies of each gene on the somatic chromosomes (chromosomes 1-22). The simplest and most common genetic variant is the single base polymorphism (SNP). Each SNP is generally found as two alleles, with the frequency of the least common allele usually 1% or greater in a population. On average, each human genome carries around 4 million SNPs. Mutations can also be single base changes in a gene but are much rarer in population. Mutations can be neutral (no effect), harmful, or beneficial – therefore, a mutation differs from a SNP in being much rarer in a population. A simple way of viewing Mendelian and complex genetic phenotypes is that a Mendelian phenotype is largely or wholly the result of a single genetic change, while complex genetic phenotypes are due to the interaction of multiple genetic changes in many genes with the environment.

A Mendelian trait segregates in simple ratios of phenotype classes due to the action of alleles at a single locus, for example, a dominant allele would result in the expression of a phenotype in heterozygotes and one homozygous class, and the ratio is 3:1. In contrast, many genetically complex traits such as height are expressed as a continuum and are measured statistically using mean and standard deviation: variance is a statistical concept that refers to the spread of values.

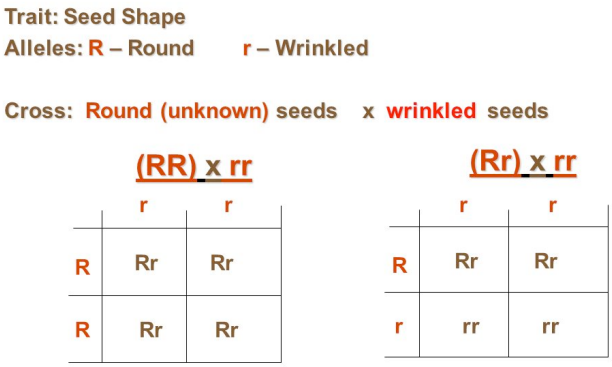


Figure 1. Example of a simple punnet square. The genotypes of the parents are shown (RR, rr, Rr), with the possible gametes of each parent forming the x- and y-axes of the square. The punnet square is a simple method for understanding the transmission of phenotypes in the context of genotypes: *Left,* all offspring are heterozygous; *Right*, a ratio of 1:1 Rr and rr

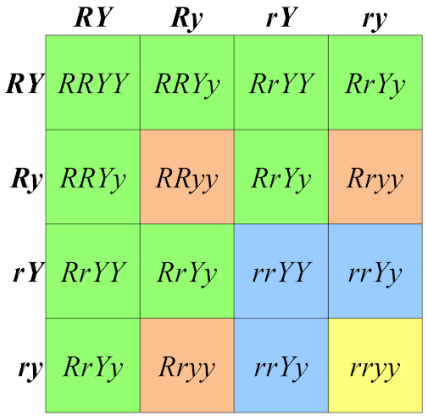


Figure 2. Segregation of two characters. R, r is rough/smooth seed coat, respectively, with R dominant; Yy is Yellow/white seed colour, respectively, with Y dominant. Similar to Figure 1, the gametes are shown on the x- and y-axes. Note that the ratio of each genotype is different from the ratio of the phenotypes (which are 9:3:3:1). The number of phenotypes is dependent upon whether there one allele is dominant or recessive, or whether there is balanced expression of each allele.

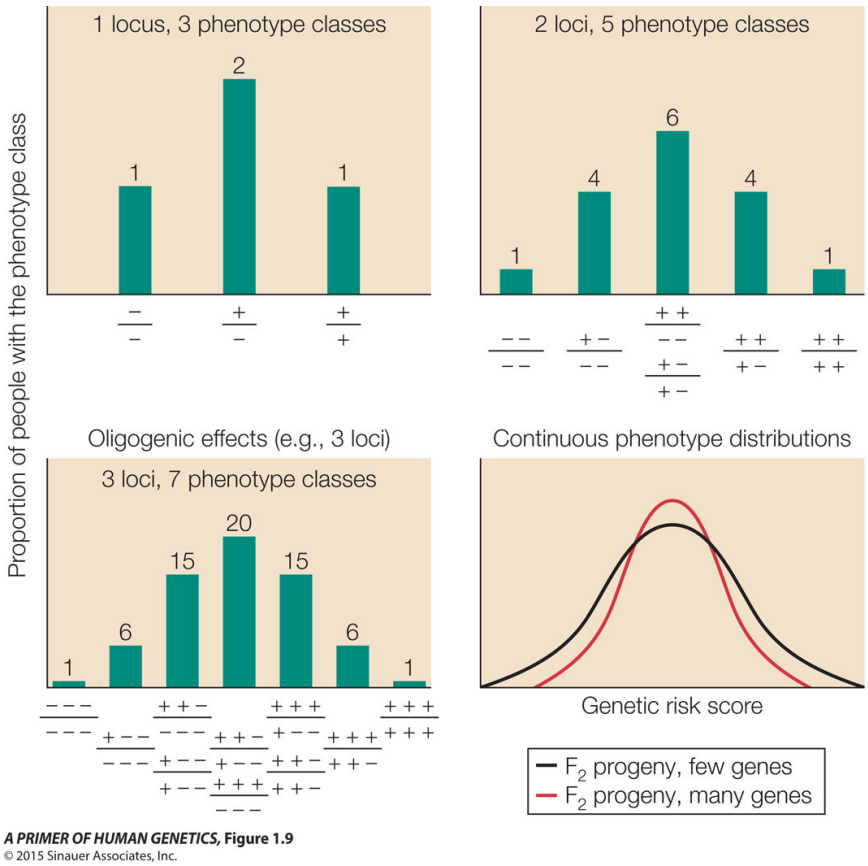


Figure 3. The genetic basis of complex phenotypes. Complex traits that are influenced by hundreds of genes and the environment can nevertheless be modelled as an extension of simple Mendelian genetics. For a single gene with two alleles (and assuming neither allele is dominant or recessive), the ratio of genotypes is 1:2:1; if two such genes are combined, the ratio is 1:4:6:4:1. As more loci are added, fewer people are at the extremes, and most people have an intermediate number of alleles that increase or decrease the trait. With the interaction of multiple loci with the environment, the categories blend into one another, resulting in a normal distribution.

As the number of loci that contribute to a specific trait or phenotype increases, the relative contribution of each allele to the phenotype tends to decrease. In a simple Mendelian disease, the presence of a single mutation will lead to expression of the disease. In this case, the mutation has high (100%) penetrance. In the case of a trait that is expressed as the result of an interaction between multiple alleles, each allele will have a small effect size, or low penetrance. If we assume that there is a “risk” allele for each SNP that contributes to a specific phenotype, individuals with the greatest number of “risk” alleles will have the greatest probability of displaying the phenotype (i.e. on the far right of the normal distribution shown in Figure 2).

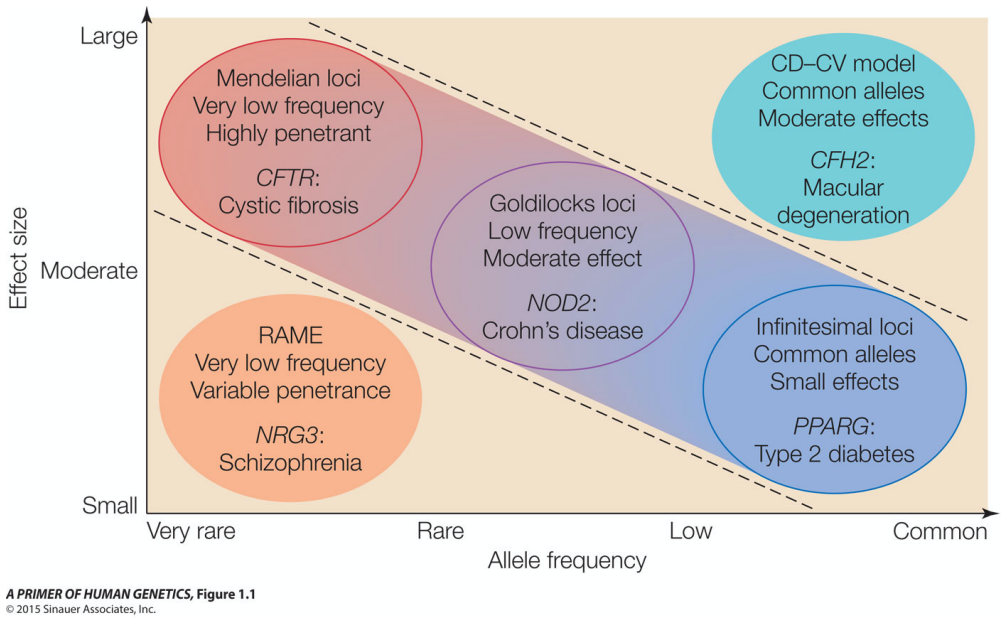


Figure 4. The Spectrum of Genetic Contributions to Disease. Most disease-associated genetic variants fall along a spectrum from very rare with large effects (Mendelian loci) to common with very small effects (the infinitesimal model of complex genetic phenotypes). The common-disease-common variant (CD-CV) model predicts may common variants with moderate to large effects, while the rare allele of major effect model predicts the opposite. The most recent genetic association studies tend to support the infinitesimal loci model.

The completion of the sequencing of the draft human genome sequence in 2000 was followed by a rapid increase in technology that has allowed the genotyping of millions of genomes, and the sequencing of many thousands of genomes.

Aims of the Online Practical Classes

1. Gain an introductory knowledge of the University of California at Santa Cruz (UCSC) Human Genome browser and database
2. Understand the key principles of reading and interpreting a scientific research paper
3. Understand how high-penetrance mutations in the gene BACH1 can lead to a Mendelian autoimmune disease and compare this with low penetrance common alleles near the same gene that also contribute to an autoimmune disease
4. Understand the basic principles of a genome-wide association study (GWAS)

Assessment: Laboratory Report 1 (25%)

The Online Practical Classes are linked to the first Assessment. A summary is provided below, but a more detailed description, including the marking rubric, can be found in your Learning Guide.

The Online Practical Classes will:

1. Introduce you to the UCSC Genome Browser
2. We will work through selected parts of two research papers (one paper in each of the two online practical classes): the first of these looks at a rare Mendelian mutation in the gene *BACH2* that is linked to immune deficiency. The second paper describes a genome-wide association study of asthma and identifies a polymorphism in the gene *BACH2* that is associated with asthma and other immune diseases.
3. Be linked to the first assessment that has a weighting of 25% and will provide training in writing and formatting a report that will be important for the second on-campus laboratory assessment that has a weighting of 35%.

Supported by online resources, we will read and discuss the content of each research paper as a class, seeking to arrive at an understanding of the rationale for the studies, the methods used, the results achieved, and the significance of the work. We will also discuss key principles of scientific writing and the formatting of figures and tables. Based on these discussions, each student will write and submit an individual report. Although discussion and collaboration is encouraged to understand the key outcomes of the research papers, the submitted work will be an individual submission. Copying of information from published sources and/or from other students is not permitted.

Online Practical Class 1

In the first online practical class, we will be exploring a rare Mendelian mutation that causes an immune deficiency, as measured by decreased numbers of lymphocytes. Lymphocytes are part of our adaptive immune system and are essential for fighting infection through the production of antibodies, as well as direct killing of infected cells (cytotoxic T-cell response).

The gene *BACH2* encodes a transcription factor, and this means we will need to have some understanding of gene structure, enhancers, and methods of studying genomes. The following figures and text provide an introduction to Online Practical Class 1.

Additional information relevant to this practical class can be found in Online Lectures 1-3

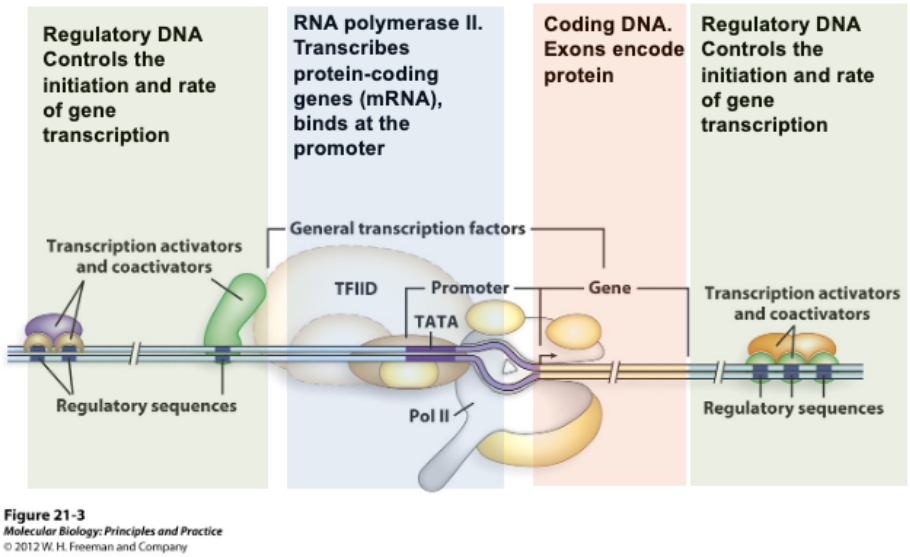


Figure 5. What is a gene? Flanking a gene are regulatory DNA sequences called that provide regulatory control over when and where (which cells) a gene will be expressed. Protein-coding genes are transcribed by RNA polymerase II that binds at a regulatory DNA sequence called the promoter that is close to the start site of transcription. Coding DNA is the exons that encode protein.

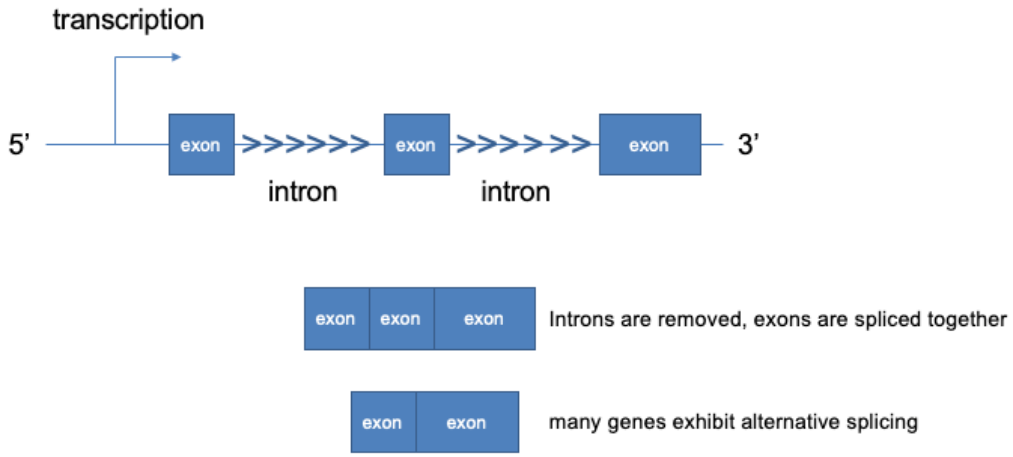


Figure 6. Transcription and splicing. RNA is transcribed in the 5' to 3' direction by RNA polymerase II. Genes are therefore given a "direction": the transcription start site of a gene is the 5' end of the gene; the final exon is at the 3' end of a gene. Regulatory DNA sequences that are found beyond the 5' end of a gene are "upstream"; regulatory DNA sequences that are found beyond the 3' end of a gene are "downstream".

As we review the two research papers for this practical class, consider where the mutations and polymorphisms are found: mutations that are associated with Mendelian disease are typically (but not always) in the coding exons of a gene and have large effects on phenotype (penetrance); common polymorphisms that are associated with common disease have a phenotypic effect through interacting with other genes and the environment and can be found in introns and regulatory DNA sequences.

In the nucleus, DNA is tightly packaged around **nucleosomes**. The complex between DNA, histones, and other protein in the nucleus is called **chromatin**. Chromatin is highly dynamic: when a gene is actively transcribed the interaction between DNA and histones is relaxed and the DNA is accessible to DNA binding proteins and RNA polymerase II to allow gene transcription to take place. Conversely, when transcription ceases or is repressed, chromatin is tightly packaged across a gene. The regulation of the packaging of chromatin and its association with gene transcription reflects reversible post-translational modifications to the **histone** subunits.

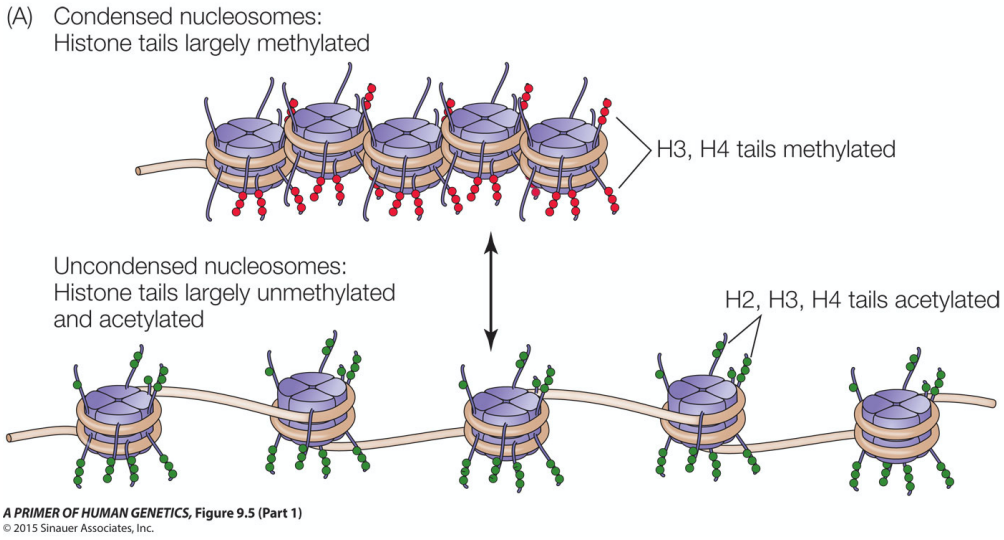


Figure 7. Histone modifications. Each histone core is an octamer of four proteins, each represented twice. The tails of those proteins can be modified at specific sites, four of which on histone subunit 3 (H3) are particularly well-characterised with respect to their general associations with transcript abundance: acetylation and H3 Lysine 4 trimethylation (H3K4me3) are associated with activation of transcription (*lower figure*); loss of acetylation and H3 Lysine 27 trimethylation (H3K27me3) are associated with repression of transcription (*upper figure*). The methylated or acetylated histone tails can be targeted with specific antibodies and used in a Chromatin Immunoprecipitation Assay (ChIP) assay (Figure 8).

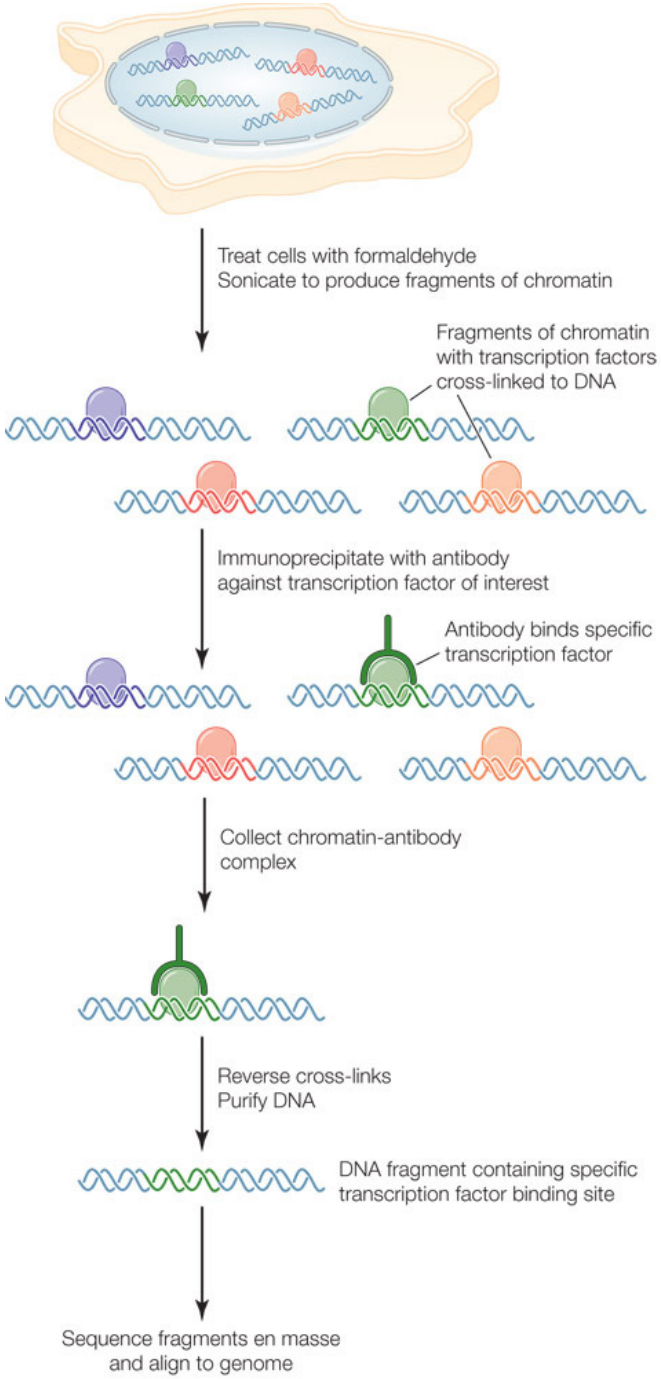


Figure 8. Chromatin Immuno-precipitation: identifying. Quantifying regions of the genome that bind transcription factors and modified histones. Areas of the genome that have an open chromatin structure are being actively transcribed. Cross-linking DNA-bound proteins to DNA, fragmenting the cross-linked products, isolating the DNA-protein fragments using specific antibodies, reverse cross-linking the protein-DNA fragments and then sequencing the DNA can identify and quantify transcribed regions of the genome.

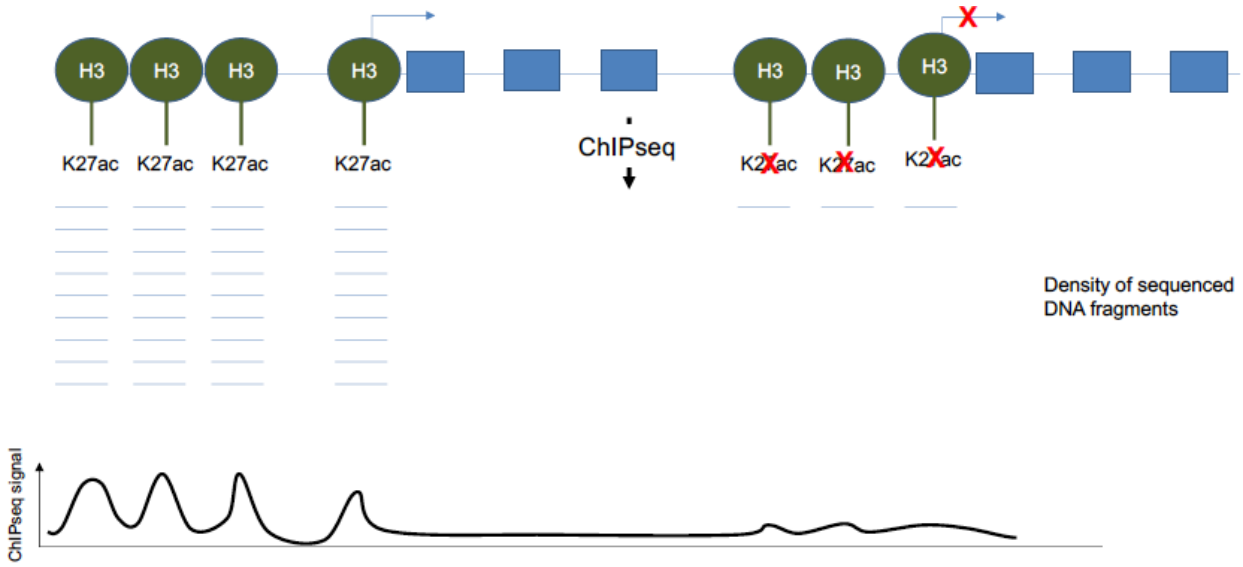
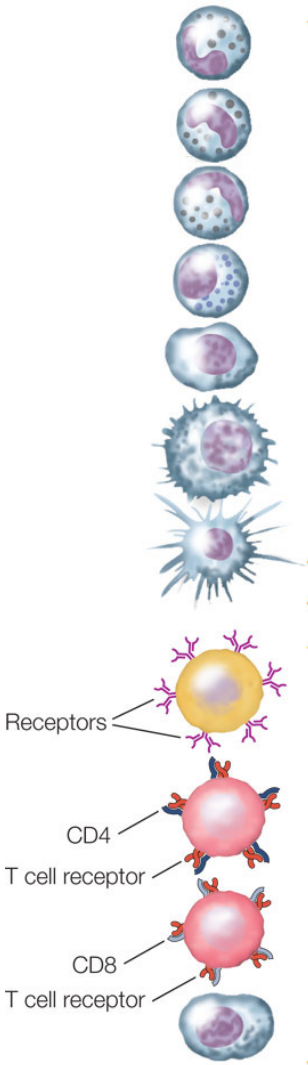


Figure 10. Simplified diagram of the principles and results of ChIPseq. In this example, the left-hand gene is being transcribed and is bound by the histone octamer that includes Histone subunit 3 (H3), acetylated on lysine residue 27 (H3K27ac). CHIPseq analysis recovers and sequences a high number of fragments that map to an enhancer and promoter for that gene (ChIPseq signal). The gene on the right is not transcribed and has lost the H3K27ac post-translation mark. No DNA fragments are recovered and there is no ChIPseq signal.



TYPE OF CELL	FUNCTION
Myeloid cells:	
Basophils (I)	Release histamine and other molecules involved in inflammation
Eosinophils (A)	Kill antibody-coated parasites
Neutrophils (I)	Stimulate inflammation
Mast cells (I)	Release histamine
Monocytes (I, A)	Develop into macrophages and dendritic cells
Macrophages (I, A)	Antigen presentation
Dendritic cells (I, A)	Present antigens to T cells
Lymphocytes:	
B cell (A)	Differentiate to form antibody-producing cells and memory cells
T helper cell (T_H) (A)	Regulate immune responses by activating or suppressing other immune cells
Cytotoxic T cell (T_C) (A)	Kill pathogen-infected cells covered in antigen and MHC-1
Natural killer cells (I, A)	Attack and lyse virus-infected or cancerous body cells

***A PRIMER OF HUMAN GENETICS*, Figure 11.1**

Figure 10. A summary of key cells of the immune system. Myeloid cells are part of the innate immune system and are the first responders to infection providing a generalised but potent immune response. The lymphocytes (but not natural killer cells) are part of the adaptive immune system and provide targeted immune memory against specific infectious agents.

Online Practical Class : Paper1 and the UCSC Genome Browser

Note: I understand these papers are highly technical, so we will focus on key experiments and outcomes. You are not required to read and understand all aspects of both papers.

Paper 1

Aim 1: The title

nature immunology

BACH2 immunodeficiency illustrates an association between super-enhancers and haploinsufficiency

Behdad Afzali[1,2,17], Juha Grönholm[3,17], Jana Vandrovcova[4,5,17], Charlotte O'Brien[5], Hong-Wei Sun[1], Ine Vanderleyden[6], Fred P Davis[1], Ahmad Khoder[5], Yu Zhang[3], Ahmed N Hegazy[7,8], Alejandro V Villarino[1], Ira W Palmer[1], Joshua Kaufman[1], Norman R Watts[1], Majid Kazemian[9], Olena Kamenyeva[3], Julia Keith[7], Anwar Sayed[5], Dalia Kasperaviciute[10], Michael Mueller[10], Jason D Hughes[11], Ivan J Fuss[3], Mohammed F Sadiyah[6], Kim Montgomery-Recht[12], Joshua McElwee[11], Nicholas P Restifo[13], Warren Strober[3], Michelle A Linterman[6], Paul T Wingfield[1], Holm H Uhlig[7,14], Rahul Roychoudhuri[6], Timothy J Aitman[5,15], Peter Kelleher[5], Michael J Lenardo[3], John J O'Shea[1], Nichola Cooper[5,18] & Arian D J Laurence[7,16,18]

The transcriptional programs that guide lymphocyte differentiation depend on the precise expression and timing of transcription factors (TFs). The TF BACH2 is essential for T and B lymphocytes and is associated with an archetypal super-enhancer (SE). Single-nucleotide variants in the *BACH2* locus are associated with several autoimmune diseases, but *BACH2* mutations that cause Mendelian monogenic primary immunodeficiency have not previously been identified. Here we describe a syndrome of BACH2-related immunodeficiency and autoimmunity (BRIDA) that results from *BACH2* haploinsufficiency. Affected subjects had lymphocyte-maturation defects that caused immunoglobulin deficiency and intestinal inflammation. The mutations disrupted protein stability by interfering with homodimerization or by causing aggregation. We observed analogous lymphocyte defects in *Bach2*-heterozygous mice. More generally, we observed that genes that cause monogenic haploinsufficient diseases were substantially enriched for TFs and SE architecture. These findings reveal a previously unrecognized feature of SE architecture in Mendelian diseases of immunity: heterozygous mutations in SE-regulated genes identified by whole-exome/genome sequencing may have greater significance than previously recognized.

A key part of our analysis of the two research papers in these online practical classes is to compare and contrast the location and penetrance of rare mutations and common polymorphisms in the context of inflammatory and autoimmune disease.

What does the title of this paper convey?

Super-enhancers

As we saw in figure 5, an **enhancer is a DNA regulatory element** that is important for the regulation of gene transcription. In complex multi-cellular organisms, a gene may be transcribed in different times in different cell types. This is particularly true of the immune system, where a gene may be expressed in several different immune cells in response to different signals (see Figure 7).

Transcription factors are proteins that bind to enhancers and promoters: there are up to 1600 genes in the human genome that encode transcription factors. Transcription of the genes encoding the transcription factors are tightly regulated to ensure transcription factors are expressed at the correct level, in the correct tissue at the correct time, and in response to the correct signal. This means that promoters and enhancers bind to multiple different transcription factors, each of which is biding to DNA in response to a specific signal. Therefore, **gene transcription is under combinatorial control**: the combination of multiple different transcription factors results in a highly specific pattern of gene transcription.

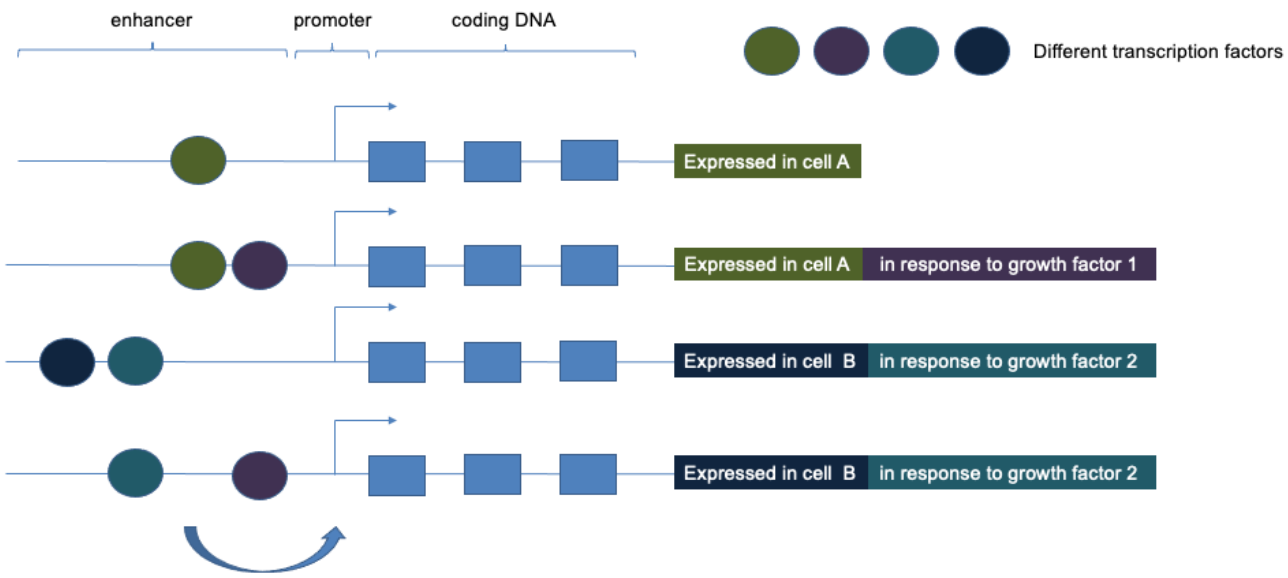


Figure 11. A highly simplified view of the combinatorial control of gene transcription. In this example, the binding of different transcription factors to an enhancer element results in different patterns of gene transcription. Each of the transcription factors is different, and this simple model of an enhancer also implies that each transcription factor binds to a different site (DNA sequence) within the enhancer.

Transcription factors bind to regulatory DNA in a sequence-specific manner. Different genes will have different combinations of short sequences (motifs) that bind different combinations of transcription factors

5–10% of genes have a complex enhancer structure consisting of multiple enhancers that are collectively described as Super-enhancers (SEs). Genes with associated SEs have a highly regulated pattern of gene expression

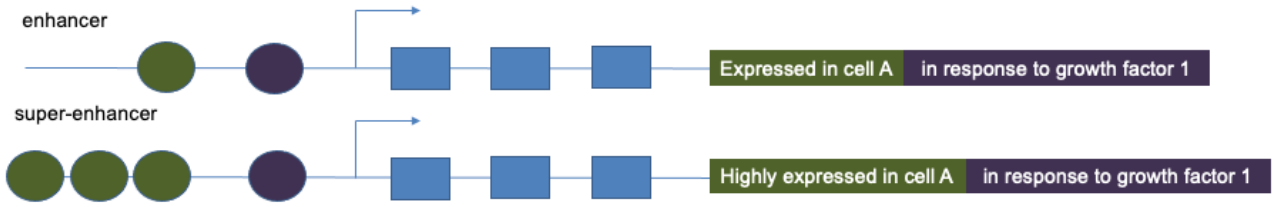


Figure 12. A simplified model of a super-enhancer. In this case, the super-enhancer has additional binding sites for the transcription factor that drives gene transcription in Cell A. This means the gene has a higher maximal expression, but can also vary over a greater range, dependent on the amount of transcription factor available in the nucleus

Now let's combine the information from figures 8 and 9 with information from figure 7: if a gene is regulated by a super-enhancer, how would the amount of transcription and histone acetylation and histone H3K4me3 methylation associated with Gene A differ to a gene not regulated by a super-enhancer?

Haploinsufficiency

We have 22 pairs of chromosomes: this means that we have 2 copies of all genes found on these chromosomes. In some disease, the loss or inactivation of one of the gene copies is sufficient to cause disease – this is haploinsuffiency.

Aim 2: UCSC Genome Browser Tutorial

https://genome.ucsc.edu/

The Human Genome Browser is a fully open-access resource for studying the Human Genome. It is used daily by scientists researching the human genome, and also has a great deal of information that is useful to clinicians working on human disease.

In this tutorial, we will take an introductory look at the genome browser, using the gene *BACH2* as an example.

You will be guided through this tutorial in the online practical class (which will be recorded).

We will:

1. Identify *BACH2* and any isoforms of *BACH2*
2. Compare the sequence of the *BACH2* gene with orthologs (homologous gene) from other organisms
3. Identify enhancers by identifying sites of histone acetylation and H3K4me3

Aim 3. What is the rationale and outcome of this study?

On the first page of the paper, read the two columns of text immediately below the abstract and the first paragraph of the second page of the paper.

Why have the authors undertaken this study?

Now read the final paragraph of the first column of text on the second page of the paper.

What is the purpose of this paragraph?

Aim 4 Results.

We will be discussing parts of the following results sections:

1. *BACH2* mutations are associated with CVID and colitis
2. *BACH2* mutations impair protein stability
3. *BACH2* mutations are not dominant negative
4. SE-regulated genes are associated with haploinsufficiency

Online Practical Class 2: Paper 2 and Genome-Wide Association Studies

Genome-wide Association Studies (GWAS)

Genomics is the study of genetic variation of the human genome in individuals and in populations. The field of genomics has only become possible with the development of high-throughput sequencing technologies that allow for the accurate and rapid sequencing a genome (***see online Lecture 3***). This has also allowed researchers and clinicians to consider the genetic basis of common but genetically complex illnesses in the human population: asthma and allergy, inflammatory diseases such as arthritis and inflammatory bowel disease, type 2 diabetes, and some mental illnesses are all examples of genetically complex illnesses.

GWAS are a **high-throughput genotyping technology** that is able to analyse differences in the frequency of common alleles of SNPs from many tens of thousands of individuals (combining multiple studies allows for the comparison of hundreds of thousands of individuals). The co-occurrence of specific alleles in a population with a specific phenotype in that population is defined as a **genetic association**. The word “association” is used to reflect the fact that this is a statistical test of populations. On average, a single human genome carries 4 million SNPs, making these genetic variants the most common variant in the human genome. While the effects of rare mutations are relatively well understood in the context of illnesses such as cancer, the phenotypic effects of SNPs and their alleles are less well understood. In general, the phenotypic effects of specific alleles of SNPs are subtle when compared to cancer-causing mutations.

Human populations have a complex history. Modern humans evolved in Africa and migrated from there into Europe, the Americas, across Asia and into the Asian Pacific region. These migrations took place over a period of at least 100,000 years. As human populations spread and settled across the world, each population needed to adapt to its local environment. This adaptation included resistance to disease, utilization of different food sources, and geographical factors such as sunlight. These adaptations are reflected in the genetic makeup of different human populations. Today, these differences can still be seen and form the basis of understanding phenotypic variation (in the context of both health and disease) in modern humans. It is for this reason that each GWAS is done for a specific population. The **Hardy-Weinberg equilibrium** test can be used to test that there is no underlying population stratification arising from the admixture of different populations that have different allele frequencies.

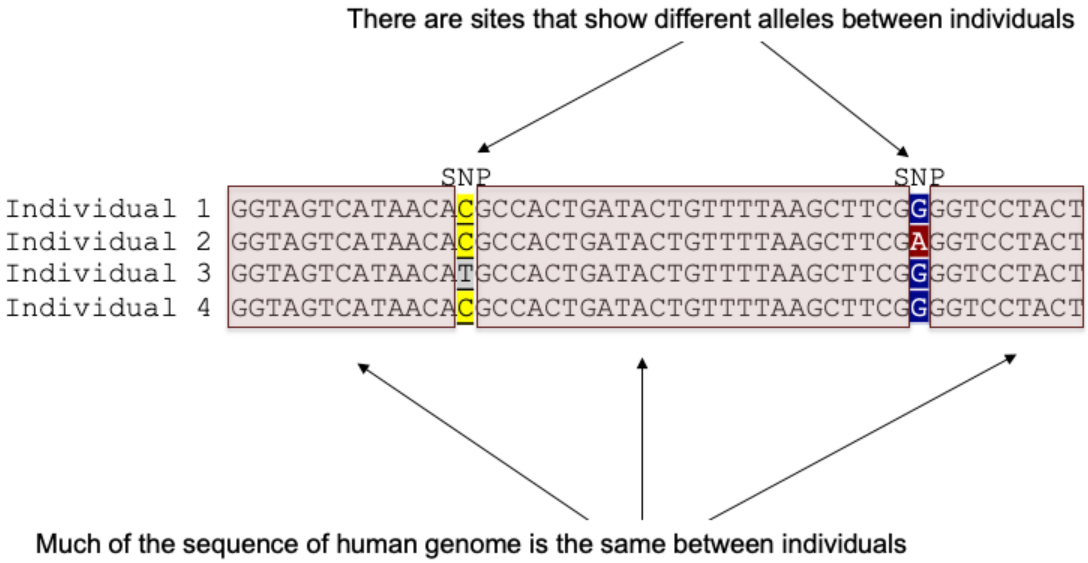


Figure 13. DNA and Genotypes. We all carry millions of genetic variants, the most common of which is the Single Nucleotide Polymorphism (SNP). Despite this, much of the 3 billion bases of the genome is identical between individuals. DNA sequencing sequences all of the bases in a genome or DNA region (locus); DNA genotyping only gathers information on the alleles in an individual genome. In a genotyping experiment the only information that would be gathered from the above example is the difference in genotypes between the four individuals (CG, CA, TG, CG).

A GWAS genotypes the entire genome of 10's to 100's of thousands of individuals in two groups of individuals: cases (those who exhibit a specific phenotype or disease) and controls (those who don't exhibit a specific phenotype or disease). The cases and controls are matched for ancestry (see the above comment relating to the **Hardy-Weinberg equilibrium**) and may also be matched for other factors such as age and diet. The rationale for matching the cases and controls is that a GWAS sets out to identify specific alleles that are more common (have a higher frequency) in the cases when compared to the controls. Any difference in allele frequency between the cases and controls is then analysed using a statistical test.

A.

Individual sample number	Genotype at a particular SNP	Disease status
S0012323	*AA*	Case
S0012324	*AA*	Case
S0012543	*AG*	Case
S0012666	*GG*	Case
S0012687	*AG*	Case
S0034301	*GG*	Control
S0034310	*GG*	Control
S0034533	*AA*	Control
S0034564	*AG*	Control
S0034662	*GG*	Control
⋮	⋮	⋮

B.

Status	*A*	*G*
Case	8680	31,320
Control	8000	32,000

Chi-square test: $p = 1.9 \times 10^{-17}$
Odds ratio (*G*:*A*) = 0.90

Figure 14. A simplified view of a GWAS. (A) 10 individuals are shown (coded for anonymity), with the result of a single SNP being shown for cases and controls; (B) The results of the GWAS, done using a total of 8680 cases and 8000 controls. The Chi-square statistical test ascertains whether there is an association between allele frequency and the presence/absence of a phenotype (disease). Note that unlike a Mendelian phenotype that is expressed in the presence of a single specific mutation, complex genetic phenotypes arise the interaction between many alleles across the genome. This means that an allele that is associated with a disease can be present in a healthy individual, but ***at the population level*** it is more common in individuals who exhibit a specific phenotype or disease. The result can also be expressed as an Odds Ratio: less than one means an allele is considered protective; greater than 1, an allele is considered a risk factor. In the above example, the G-allele is protective. In the cases, the frequency of the G-allele is 0.78 while in the controls the frequency of the G-allele is 0.8. Verifying that such small differences are statistically significant can only be done by recruiting very large sample numbers.

Aim 1. Complete calculations for a simple GWAS example.

The aims of this part of the class are:

1. Test population structure using the Hardy-Weinberg equilibrium
2. Use the Chi-square test to test for association
3. Calculate the Odds Ratios for each SNP tested
4. Discussion on the results

The spreadsheet for this class can be found in the Online Practical Classes folder on vUWS.

The spreadsheet has the following tabs:

1. SNP GWAS Details: This includes the SNP identifying rs number, as well as genotype frequencies for each of the genotypes for each SNP
2. Hardy-Weinberg Test
3. ChiSquare Test: This will test for the association between each allele and inflammatory disease by producing a p value
4. Calculate Odds Ratio: This sheet will allow you to calculate Odds Ratios by entering the allele counts for each SNP

Aim 2. Paper 2. The title, abstract, and introduction.

ARTICLES
https://doi.org/10.1038/s41588-017-0014-7
nature genetics

Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks

We examined common variation in asthma risk by conducting a meta-analysis of worldwide asthma genome-wide association studies (23,948 asthma cases, 118,538 controls) of individuals from ethnically diverse populations. We identified five new asthma loci, found two new associations at two known asthma loci, established asthma associations at two loci previously implicated in the comorbidity of asthma plus hay fever, and confirmed nine known loci. Investigation of pleiotropy showed large overlaps in genetic variants with autoimmune and inflammatory diseases. The enrichment in enhancer marks at asthma risk loci, especially in immune cells, suggested a major role of these loci in the regulation of immunologically related mechanisms.

What does the title convey?

What does the abstract tell you about the study?

Read the first paragraph below the introduction.

- What is the rationale for conducting this study?
- What does it tell you about the genetic basis of asthma?
- How was the study done?
- What are the outcomes of the study?

Aim 3. Discussion of Figure 1 and Table 1.

Figure 1 is a graphical summary of the GWAS, and is known as a “Manhattan Plot” (after the profile of the New York skyline). Table 1 presents key numerical information about the outcomes of the study and will provide us with insight about the structure of common complex genetic disease.

Aim 4. Location of Associated SNPs, with a focus on *BACH2.*

- We will use the UCSC genome browser to locate the SNP rs2325291, with reference to the location of features such as histone H3 acetylation, introns, and exons
- Consider this analysis in the context of information presented in Table 4, and in the results sections “Overlap of loci associated with asthma and other phenotypes” and “Enrichment of asthma risk loci in epigenetic marks”
- What is the conclusion reached by this study?

Conclusion of Online Practical Classes 1 and 2.

Bringing it all together: a discussion of what you have learnt about the genetics of Mendelian and complex genetic phenotypes and how to prepare for the first assessment.