Glossary of Terms

GLOSSARY OF GENETIC TERMS:

(prepared by Gurdeep Sagoo, University of Sheffield, UK)

N.B. Some of the definitions below assume that the organism of interest is diploid.

Adenine (A): purine base that forms a pair with thymine in DNA and uracil in RNA.

Admixture: arises when two previously isolated populations begin interbreeding.

Allele: one of the possible forms of a gene at a given locus. Depending on the technology used to type the gene, it may be that not all DNA sequence variants are recognised as distinct alleles.

Allele frequency: often used to mean the relative frequency (i.e. proportion) of an allele in a sample or population.

Allelic association: the non-independence, within a given population, of a gamete's alleles at different loci. Also commonly (and misleadingly) referred to as *linkage disequilibrium.*

Alpha helix: a helical (usually right-handed) arrangement that can be adopted by a polypeptide chain; a common type of protein secondary structure.

Amino acid: the basic building block of proteins. There are 20 naturally occurring amino acids in animals which when linked by peptide bonds form polypeptide chains.

Aneuploid cells: do not have the normal number of chromosomes.

Antisense strand: the DNA strand complementary to the coding strand, determined by the covalent bonding of A with T and C with G.

Ascertainment: the strategy by which individuals are identified, selected, and recruited for participation in a study.

Autosome: A chromosome other than the sex chromosomes. Humans have 22 pairs of autosomes plus 2 sex chromosomes.

Backcross: A linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed back to one of the inbred parental strains.

Bacterial Artificial Chromosome (BAC): a vector used to clone a large segment of DNA (100–200 Kb) in bacteria resulting in many copies.

Base: (abbreviated term for a purine or pyrimidine in the context of nucleic acids), a cyclic chemical compound containing nitrogen that is linked to either a deoxyribose (DNA) or a ribose (RNA).

Base pair (bp): a pair of bases that occur opposite each other (one in each strand) in double stranded DNA/RNA. In DNA adenine base pairs with thymine and cytosine with guanine. RNA is the same except that uracil takes the place of thymine.

Bayesian: A statistical school of thought that, in contrast with the frequentist school, holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Bayes Theorem allows one to compute the posterior distribution for an unknown from the observed data and its assumed prior distribution.

Beta-sheet: is a (hydrogen-bonded) sheet arrangement which can be adopted by a polypeptide chain; a common type of protein secondary structure.

centiMorgan (cM): measure of genetic distance. Two loci separated by 1 cM have an average of 1 recombination between them every 100 meioses. Because of the variability in recombination rates, genetic distance differs from physical distance, measured in base pairs. Genetic distance differs between male and female meioses; an average over the sexes is usually used.

Centromere: the region where the two sister chromatids join, separating the short (p) arm of the chromosome from the long (q) arm.

Chiasma: the visible structure formed between paired homologous chromosomes (non-sister chromatids) in meiosis.

Chromatid: a single strand of the (duplicated) chromosome, containing a double-stranded DNA molecule.

Chromatin: the material composed of DNA and chromosomal proteins that makes up chromosomes. Comes in two types, euchromatin and heterochromatin.

Chromosome: the self-replicating threadlike structure found in cells. Chromosomes, which at certain stages of meiosis and mitosis consist of two identical sister chromatids, joined at the centromere, and carry the genetic information encoded in the DNA sequence.

cis-Acting: regulatory elements and eQTL whose DNA sequence directly influences transcription. The physical location for cis-acting elements will be in or near the gene or genes they regulate.

Clones: genetically engineered identical cells/sequences.

Co-dominance: both alleles contribute to the phenotype, in contrast with recessive or dominant alleles.

Codon: a nucleotide triplet that encodes an amino acid or a termination signal.

Common disease common variant (CDCV) hypothesis: The hypothesis that many genetic variants underlying complex diseases are common, and hence susceptible to detection using current population association study designs. An alternative possibility is that genetic contributions to the causation of complex diseases arise from many variants, all of which are rare.

complementary DNA (cDNA): DNA that is synthesised from a messenger RNA template using the reverse transcriptase enzyme.

Contig: a group of contiguous overlapping cloned DNA sequences.

Cytosine (C): pyrimidine base that forms a pair with guanosine in DNA.

Degrees of freedom (df): This term is used in different senses both within statistics and in other fields. It can often be interpreted as the number of values that can be defined arbitrarily in the specification of a system; for example, the number of coefficients in a regression model. Frequently it suffices to regard df as a parameter used to define certain probability distributions.

Deoxyribonucleic acid (DNA): polymer made up of deoxyribonucleotides linked together by phosphodiester bonds.

Deoxyribose: the sugar compound found in DNA.

Diploid: has two versions of each autosome, one inherited from the father and one from the mother. Compare with haploid.

Dizygotic twins: twins derived from two separate eggs and sperm. These individuals are genetically equivalent to full sibs.

DNA methylation: the addition of a methyl group to DNA. In mammals this occurs at the C-5 position of cytosine, almost exclusively at CpG dinucleotides.

DNA microarray: small slide or 'chip' used to simultaneously measure the quantity of large numbers of different mRNA gene transcripts present in cell or tissue samples.

Depending on the technology used, measurements may either be absolute or relative to the quantities in a second sample.

Dominant allele: results in the same phenotype irrespective of the other allele at the locus.

Effective population size: The size of a theoretical population that best approximates a given natural population under an assumed model. The criterion for assessing the 'best' approximation can vary, but is often some measure of total genetic variation.

Enzyme: a protein that controls the rate of a biochemical reaction.

Epigenetics: the transmission of information on gene expression to daughter cells at cell division.

Epistasis: the physiological interaction between different genes such that one gene alters the effects of other genes.

Epitope: the part of an antigen that the antibody interacts with.

Eukaryote: organism whose cells include a membrane-bound nucleus. Compare with prokaryote.

Exons: parts of a gene that are transcribed into RNA and remain in the mature RNA product after splicing. An exon may code for a specific part of the final protein.

Expression Quantitative Trait Locus (eQTL): a locus influencing the expression of one or more genes.

Fixation: occurs when a locus which was previously polymorphic in a population becomes monomorphic because all but one allele has been lost through genetic drift.

Frequentist: the name for the school of statistical thought in which support for a hypothesis or parameter value is assessed using the probability of the observed data (or more 'extreme' datasets) given the hypothesis or value. Usually contrasted with Bayesian.

Gamete: a sex cell, sperm in males, egg in females. Two haploid gametes fuse to form a diploid zygote.

Gene: a segment (not necessarily contiguous) of DNA that codes for a protein or functional RNA.

Gene expression: the process by which coding DNA sequences are converted into functional elements in a cell.

Genealogy: the ancestral relationships among a sample of homologous genes drawn from different individuals, which can be represented by a tree. Also sometimes used in

place of pedigree, the ancestral relationships among a set of individuals, which can be represented by a graph.

Genetic drift: the changes in allele frequencies that occur over time due to the randomness inherent in reproductive success.

Genome: all the genetic material of an organism.

Genotype: the (unordered) allele pair(s) carried by an individual at one or more loci. A multilocus genotype is equivalent to the individual's two haplotypes without the phase information.

Guanine (G): purine base that forms a pair with cytosine in DNA.

Haemoglobin: is the red oxygen-carrying pigment of the blood, made up of two pairs of polypeptide chains called globins (2α and 2β subunits).

Haploid: has a single version of each chromosome.

Haplotype: the alleles at different loci on a chromosome. An individual's two haplotypes imply the genotype; the converse is not true, but in the presence of strong linkage disequilibrium haplotypes may be inferred from genotype with few errors.

Hardy-Weinberg disequilibrium: the non-independence within a population of an individual's two alleles at a locus; can arise due to inbreeding or selection for example. Compare with linkage disequilibrium.

Heritability: the proportion of the phenotypic variation in the population that can be attributed to underlying genetic variation.

Heterozygosity: the proportion of individuals in a population that are heterozygotes at a locus. Also sometimes used as short for expected heterozygosity under random mating, which equals the probability that two homologous genes drawn at random from a population are not the same allele.

Heterozygote: a single-locus genotype consisting of two different alleles.

HIV (Human Immunodeficiency Virus): a virus that causes acquired immune deficiency syndrome (AIDS) which destroys the body's ability to fight infection.

Homology: similarities between sequences that arise because of shared evolutionary history (descent from a common ancestral sequence). Homology of different genes within a genome is called paralogous while that between the genomes of different species is called orthologous.

Homozygote: a single-locus genotype consisting of two versions of the same allele.

Hybrid: the offspring of a cross between parents of different genetic types or different species.

Hybridization: the base pairing of a single stranded DNA or RNA sequence, usually labelled, to its complementary sequence.

ibd: identity by descent; two genes are ibd if they have descended without mutation from an ancestral gene.

Inbred lines: derived and maintained by repeated selfing or brother–sister mating, these individuals are homozygous at essentially every locus.

Inbreeding: either the mating of related individuals (e.g. cousins) or a system of mate selection in which mates from the same geographic area or social group for example are preferred. Inbreeding results in an increase in homozygosity and hence an increase in the prevalence of recessive traits.

Intercross: A linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed or selfed. This design is also sometimes referred to as an *F2 design* because the resulting individuals are known as F2s.

Intron: non-coding DNA sequence separating the exons of a gene. Introns are initially transcribed into messenger RNA but are subsequently spliced out.

Karyotype: the number and structure of an individual's chromosomes.

Kilobase (Kb): 1000 base pairs.

Linkage: two genes are said to be linked if they are located close together on the same chromosome. The alleles at linked genes tend to be co-inherited more often than those at unlinked genes because of the reduced opportunity for an intervening recombination.

Linkage disequilibrium (LD): the non-independence within a population of a gamete's alleles at different loci; can arise due to linkage, population stratification, or selection. The term is misleading and 'gametic phase disequilibrium' is sometimes preferred. Various measures of linkage disequilibrium exist.

Locus (pl. Loci): the position of a gene on a chromosome.

LOD score: a likelihood ratio statistic used to infer whether two loci reside close to one another on a chromosome and are therefore inherited together. A LOD score of 3 or more is generally thought to indicate that the two loci are close together and therefore linked.

Marker gene: a polymorphic gene of known location which can be readily typed; used for example in genetic mapping.

Megabase (Mb): 1000 kilobases = 1,000,000 base pairs.

Meiosis: the process by which (haploid) gametes are formed from (diploid) somatic cells.

messenger RNA (mRNA): the RNA sequence that acts as the template for protein synthesis.

Microarray: see DNA microarray above.

Microsatellite DNA: small stretches of DNA (usually 1–4 bp) tandemly repeated. Microsatellite loci are often highly polymorphic, and alleles can be distinguished by length, making them useful as marker loci.

Mitosis: the process by which a somatic cell is replaced by two daughter somatic cells.

Monomorphic: a locus at which only one allele arises in the sample or population.

Monozygotic twins: genetically identical individuals derived from a single fertilized egg.

Morgan: 100 centiMorgans.

mtDNA: the genetic material of the mitochondria which consists of a circular DNA duplex inherited maternally.

Mutation: a process that changes an allele.

Negative selection: removal of deleterious mutations by natural selection. Also known as *purifying selection.*

Neutral: not subject to selection.

Neutral evolution: evolution of alleles with nearly zero selective coefficient. When $|Ns| << 1$, where N is the population size and s is the selective coefficient, the fate of the allele is mainly determined by random genetic drift rather than natural selection.

Non-Coding RNA (ncRNA): RNA segments that are coded for in the genome, but not translated into protein product. Composed of many classes, the complete range of functions of these molecules has yet to be characterised, but they have been shown to affect the rate of transcription and transcript degradation.

Nonsynonymous substitution: Nucleotide substitution in a protein-coding gene that alters the encoded amino acid.

Nucleoside: a base attached to a sugar, either ribose or deoxyribose.

Nucleotide: the structural units with which DNA and RNA are formed. Nucleotides consist of a base attached to a five-carbon sugar and mono-, di-, or tri-phosphate.

Nucleotide substitution: the replacement of one nucleotide by another during evolution. Substitution is generally considered to be the product of both mutation and selection.

Oligonucleotide: a short sequence of single-stranded DNA or RNA, often used as a probe for detecting the complementary DNA or RNA.

Open Reading Frame (ORF): a long sequence of DNA with an initiation codon at the 5′-end and no termination codon except for one at the 3′-end.

PCR (polymerase chain reaction): a laboratory process by which a specific, short, DNA sequence is amplified many times.

Pedigree: a diagram showing the relationship of each family member and the heredity of a particular trait through several generations of a family.

Penetrance: the probability that a particular phenotype is observed in individuals with a given genotype. Penetrance can vary with environment and the alleles at other loci for example.

Peptide bond: linkages between amino acids occur through a covalent peptide bond joining the C terminal of one amino acid to the N terminal of the next (with loss of a water molecule).

Phase (of linked markers): the relationship (either coupling or repulsion) between alleles at two linked loci. The two alleles at the linked loci are said to be in coupling if they are present on the same physical chromosome or in repulsion if they are present on different parental homologs.

Phenotype: the observed characteristic under study, may be quantitative (i.e. continuous) such as height, or binary (e.g. disease/no disease), or ordered categorical (e.g. mild/moderate/severe).

Pleiotropy: is the effect of a gene on several different traits.

Polygenic: influenced by more than one gene.

Polymorphic: a locus that is not monomorphic. Usually a stricter criterion is imposed: a locus is polymorphic only if no allele has frequency over 99 %.

Polynucleotide: a polymer of either DNA or RNA nucleotides.

Polypeptide: is a long chain of amino acids joined together by peptide bonds.

Polypeptide chain: A series of amino acids linked by peptide bonds. Short chains are sometimes referred to as oligopeptides or simply peptides.

Polytene: refers to the giant chromosomes that are generated by the successive replication of chromosome pairs without the nuclear division, thus several chromosome sets are joined together.

Population stratification (or population structure): Refers to a situation in which the population of interest can be divided into strata such that an individual tends to be more closely related to others within the same stratum than to other individuals.

Positive selection: fixation, by natural selection, of an advantageous allele with a positive selective coefficient. Also known as *Darwinian selection.*

Proband: an individual through whom a family is ascertained, typically by their phenotype.

Prokaryote: organism whose cells have no nucleus.

Promoter: located upstream of the gene, the promoter allows the binding of RNA polymerase which initiates transcription of the gene.

Protein: a large, complex, molecule made up of one or more chains of amino acids.

Pseudogene: a DNA sequence that is either an imperfect, non-functioning, copy of a gene, or a former gene which no longer functions due to mutations.

Purine and Pyrimidine: are particular kinds of nitrogen containing heterocyclic rings.

Purine: adenine or guanine.

Pyrimidine: cytosine, thymine, or uracil.

QTL (Quantitative Trait Locus): a locus influencing a continuously varying phenotype.

Radiation hybrid: a cell line, usually rodent, that has incorporated fragments of foreign chromosomes that have been broken by irradiation. They are used in physical mapping.

Recessive allele: has no effect on phenotype except when present in homozygote form.

Recombination: the formation of new haplotypes by physical exchange between two homologous chromosomes during meiosis.

Restriction enzyme: recognises specific nucleotide sequences in double-stranded DNA and cuts at a specified position with respect to the sequence.

Restriction fragment: a DNA fragment produced by a restriction enzyme.

Restriction site: a 4–8 bp DNA sequence (usually palindromic) that is recognised by a restriction enzyme.

Retrovirus: an RNA virus whose replication depends on a reverse transcriptase function, allowing the formation of a cDNA copy that can be stably inserted into the host chromosome.

Ribonucleic acid (RNA): polymer made up of ribonucleotides that are linked together by phosphodiester bonds.

Ribosome: a cytoplasmic organelle, consisting of RNA and protein, that is involved in the translation of messenger RNA into proteins.

Ribosomal RNA (rRNA): the RNA molecules contained in ribosomes.

Selection: a process such that expected allele frequencies do not remain constant, in contrast with genetic drift. Alleles that convey an advantage to the organism in its current environment tend to become more frequent in the population (positive, or adaptive, selection), while deleterious alleles become less frequent. Under stabilising (or balancing) selection, allele frequencies tend towards a stable, intermediate value.

Sense strand: the DNA strand in the direction of coding.

Sex-linked: a trait influenced by a gene located on a sex (X or Y) chromosome.

Single nucleotide polymorphism (SNP): a polymorphism consisting of a single nucleotide.

Sister chromatids: two chromatids that are copies of the same chromosome. Non-sister chromatids are different but homologous.

Somatic cell: a non-sex cell.

Synonymous substitution: Nucleotide substitution in a protein-coding gene that does not alter the encoded amino acid.

TATA box: a conserved sequence (TATAAAA) found about 25–30 bp upstream from the start of transcription site in most but not all genes.

Thymine (T): pyrimidine base that forms a pair with adenine in DNA.

trans-Acting: eQTL whose DNA sequence influences gene expression through its gene product. These regulatory elements are often coded for at loci far from or unlinked to the genes they regulate.

Transcription: the synthesis of a single-stranded RNA version of a DNA sequence.

Transition: a mutation that changes either one purine base to the other, or one pyrimidine base to the other.

Translation: the process whereby messenger RNA is 'read' by transfer RNA and its corresponding polypeptide chain synthesized.

Transposon: a genetic element that can move over generations from one genomic location to another.

Transversion: a mutation that changes a purine base to a pyrimidine, or vice-versa.

Uracil (U): pyrimidine base in RNA that takes the place of thymine in DNA, also forming a pair with adenine.

Wild-type: the common, or standard, allele/genotype/phenotype in a population.

Yeast artificial chromosome (YAC): a cloning vector able to carry large (e.g. one megabase) inserts of DNA and replicate in yeast cells.

Zygote: an egg cell that has been fertilized by a sperm cell.