1 Molecular Genetics

Here we define some fundamental molecular biology from a genetics standpoint, building first from the components of DNA and moving up towards the nature of inheritance. For those who already have a firm understanding of these concepts, this chapter may be skipped. Otherwise, the aim here is to define the terminology and outline the principles that will be used in the rest of the thesis.

1.1 DNA

Each organism is defined by the data encoded within DNA(deoxyribonucleic acid), a long molecule made of up sub-units of *nucleotides* that constitutes the *genome* of the organism. The order of these nucleotides form an ordered sequence specific to that individual, and ultimately encoded the genetic information that makes that individual unique.

DNA is packed into *chromosomes*, two long strands of chained nucleotides bound together by hydrogen bonds between nucleotide pairs on opposing strands. Nucleotides consist of a molecule made up of a phosphate group, 2-deoxyribose, and a *nucleobase* (or base). There are four possible bases: adenine, cytosine, guanine, or thymine (or A, C, T, G). All adenine bases are paired with thymine, and all cytosine bases are paired with guanine.
These form AT and GC pairs across the two strands forming what are known as *base pairs* (or bp). The sugar and phosphate groups alternate in a chain along the side of DNA, lending a backbone structure on the outer edge of the bases that forms the eponymous helical shape known as *the double helix*. In humans, there are 3.3 billion base pairs of DNA.

1.1.1 Polarity

Every strand of DNA has a defined polarity, for the sake of a consistent reading orientation. The two strands of the double helix lie in opposite directions of one another, such that the starting end of one strand is the finishing end of the other.

The starting end is referred to as 5’ (five prime), and the finishing end is known as a 3’. These numbers are named in accordance to the clockwise ordering of the carbon atoms in the deoxyribose molecule.

<Diagram of the molecule showing 1’ to 5’>

Caption: The 5’ end is terminated with a phosphate, and the 3’ end with a hydroxyl sugar group.

For consistency, biologists follow the convention of defining the 5’ to 3’ orientation as the *forward* (or *‘sense*’) strand. The 3’ to 5’ orientation is the *reverse* (or ‘*antisense*’) strand.

1.1.2 Chromosomes

Chromosomes hold packaged DNA in the form of *chromatin*, as well as specialised proteins that perform chromosome-specific tasks such as a reading, replicating, and repairing the DNA. There are many chromosomes within a cell; in a human there are 23 distinct chromosomes within a *haploid* cell such as gametes (sperm and unfertilized eggs), and 23 distinct *pairs* of chromosomes in a *diploid* cell which contain an individual’s entire genome.

Chromosomes the end stages of cell division the structure of chromosomes can be clearly seen, with two arms extending out from each strand, with the chromosomes bound together in the middle. The *p* and the *q* arm of each chromosome denotes the small and larger arm respectively. These arms are split into bands as defined by the density of the chromatin in the region, and a specific locus can be specified by, e.g. 16p5.2, which denotes that the locus of interest lies on chromosome 16, on the p-arm, in band 5, sub-band 2.

The middle section of the chromosome in known as the *centromere* and is purely structural; it does not encode any data, but holds pairs of chromosomes together. The ends are capped with *telomeres* which act as disposable buffers, and are repeatedly truncated during cell division.

Pairs of chromosomes are known as *homologs*, and are split into two groups: *autosomal* chromosomes, and *allosomal* chromosomes (or 'sex-specific' chromosomes).

The autosomes are the most common with 22 pairs of chromosomes within the group. The last remaining pair are therefore of the allosomal variety, and determines the sex of an individual during sexual reproduction and are typically what are known as X and Y chromosomes. Females have a homologous pair of X-X chromosomes, whereas males have an X and Y.

Each pair of chromosomes has a complete set of genes, assuming no duplications or deletions. These two copies of each gene are referred as *alleles* of the gene, though the term itself can describe any locus between chromosome pairs.

1.1.3 Genes and Function

Genes are the parts of DNA that are mostly responsible for generating a *phenotype,* the observable physical characteristics of an organism. The sections of DNA that comprise gene data are known as the *coding* regions or *exons*, and are delimited by *codons* which are triplet groups of nucleobases that together form a code to be read by *RNA polymerase* enzymes, and *ribosome* proteins that translate it for protein synthesis in an uninterrupted run of codons known as the *open reading frame*.

Coding regions are sparsely populated in the human genome, occupying 1-2% of the sequence. The remainder consists of either unknown or regulatory regions responsible for gene expression. Regions between genes are known as *intergenic* and may play some regulatory roles, but are not typically bound to any particular gene.

Genes are made up of four main components: 5'UTR, exons, introns, splice sites, and 3'UTR.

UTR (or 'untranslated regions') cap the gene extremes with a 5' upstream end containing binding sites to initiate *transcription* via a *start codon*, and a 3' downstream end with sites to terminate the

process via a *stop codon*. The transcription process essentially clones a given DNA segment into a precursor RNA molecule called messenger RNA (or *mRNA*), which is similar to the DNA molecule except that the sugar is a ribose instead of a deoxyribose, and thymine is replaced with a base called *uracil* (U).

Exons constitute the coding regions of the gene, alternated by introns which are non-coding. Splice sites flank the exon-intron boundaries and are responsible for the *splicing* process which binds coding (exon) regions together during transcription for the open reading frame into spliced mRNA. The splice mRNA is then *translated* into a protein by a ribosome, with several different proteins being made from the same DNA template due to splice variations. For reference, there are currently approximately 60,000 genes in the human genome[1] that give rise to 45 million proteins[2].

Proteins are a product of their *amino-acid* sequence derived from DNA codons, as well as the emergent properties from the contorted three dimensional structure they exhibit due to their folded arrangement. There are 20 different amino-acid bases that are one-to-many mapped from 64 different DNA codons. This redundancy has some interesting caveats, the most notable being that there is only one start codon (ATG) and three stop codons (TAA, TGA, TAG).

1.2 Heredity

During sexual reproduction, chromosomes from both the mother and father are assorted and fused together to create new chromosomes that are passed on into their offspring. This 'reshuffling' of DNA is what creates such variation in organisms, such that the chance of any two individuals of the same organism being clones of each other are extremely slim.

1 As of Human Genome version GRCh38, reference Gene database.

2 As of UniProtKB database release version June 2016, 75% of which are predicted.

Chromosomes are assorted during the process of *meiosis*, where during reproduction a diploid cell will split into four distinct haploid cells (gametes). The maternal and paternal gametes then merge to form a diploid *zygote* which contains DNA from both parents.

1.2.1 Meiosis

Meiosis is split into three phases: *meiotic S, meiosis I*, and *meiosis II.*

In the S-phase, homologs from the chromosomes of each parent are replicated. This results in a cell that has twice the number of chromosomes.

1.2.1.1 Meiosis I

The meiosis I phase is by far the longest phase of meiosis, and is split into nine stages.

First, homologs from each parent pair up and exchange the DNA in their sister chromatids (**Prophase I**). Sister chromatids of each chromosome then detach into thin threads (**Leptotene**), and the chromosomes then align into homologous chromosome pairs that snap together outwards in a zipper-like fashion from the centromere, forming a *bivalent* (**Zygotene**).

The chromatids now exchange homologous regions of DNA to non-sister chromatids, recombining their data in a non-deleterious fashion (**Pachytene**). Homologous chromosomes then separate and uncoil to allow limited transcription, and all bivalents condense such that points of recombination entangle (**Diplotene**). A *meiotic spindle* becomes pronounced between sister chromatids. Genetic content then migrates to the two poles of the cell (**Diakinesis**) .

The bivalents migrate to a *metaphase plate* plane and are randomly oriented to one another as a precursor to the independent assortment of the chromosomes (**Metaphase I**). Homologous chromosomes then move to opposite poles, with the sister chromatids remaining intact as the

homologs are separated (**Anaphase I**). The meiotic spindle disappears and cell de-condensates and lengthens, completing cell division (**Telophase I**).

The two resultant daughter cells have half the number of original chromosomes, but each chromosome is made up of a pair of chromatids.

1.2.1.2 Meiosis II

Meiosis II can be seen as a simplified version of the cell division in the previous phase, as many of the steps are repeated.

Chromatids once again shorten and thicken (**Prophase II**), and as before the meiotic spindle surfaces dragging material to the poles (**Metaphase II**). Sister chromatids then segregate towards the poles and become sister chromosomes (**Anaphase II**), and finally the spindle is dissembled and the chromosomes lengthened (**Telophase II**).

This results in four haploid gametes from the two daughter cells in meiosis I.

1.2.2 Recombination

Recombination occurs during meiosis I, when the bivalents are formed from the parental homologs bound by connections known as *chiasmata*. The process involves splitting and recombining segments of DNA segments across sister chromatids at the chiasmata. This results in exchanged genetic material at a specific point on a pair of chromosomes known as a *crossover*, with a single crossover event occurring per meioses.

The chance of a crossover event occurring is based on can be determined between two loci. This probability (or *recombination frequency*) is in the range [0,0.5] increasing with the distance between the loci.

For example, two adjacent loci in close proximity to one another will have a near zero chance of a recombination event occurring. However, if the loci were on different chromosomes, then it's safe to assume that the two loci will segregate independently, with a probability tending towards 0.5.

Any higher would assume that there was *linkage disequilibrium* between the two locations and that the chance of a recombination is biased. (XXX: Robert: - is this statement correct?)

Recombination frequencies are not uniform throughout a chromosome, as the nature and density of the underlying chromatin influences the chance across a given band. The frequencies are also sex-specific, since women are more prone to crossover events than men.

Haldane's model of recombination modelled crossovers as a Poisson[3] distribution, defining the *Morgan,* a unit of genetic distance, that details the expected number of crossovers between two loci. The Morgan is defined such that there is (on average) 1 expected crossover event occurring at a distance 1 Morgan. In practical circumstances, the centiMorgan sub-unit is used to refer to 0.01 expected crossovers occurring at a distance of 1cM.

$$\theta = ½\left(1 - e^{-2d}\right) \tag{1}$$

Note that Haldane's model does not model *crossover interference*, where the act of a crossover prevents the nearby act of another crossover from happening.

It is clear that recombination events are instrumental to our understanding of how offspring inherit traits from their parents, specifically where the crossovers occur so that we can trace the flow of data throughout the generations.

3See Appendix section XXX

In order to do so, we must first map chromosomes so that we can identify loci closely related to the inherited traits in question.

1.3 Molecular Maps

Historically, mapping traits to a locus was performed through extensive breeding experiments and then tabulating the number of different traits that appeared. These traits would then act as sign-posts for where the trait or, more typically, the disease phenotype would lie. The resolution of these methods were limited however, with the disease locus being only distinguishable between different chromosomes.

Advances in the field now make use of numerous flags or *markers*, evenly-spaced across the chromosome such that the phenotype can be located by any two flanking markers that surround the region of interest. The effectiveness of their usage is determined by how well they conform to following principles:

- **Known locus** – The trait of interest lies in an unknown location, but it's position can be inferred relative to a marker with a known position.
- **Polymorphic** – The marker must denote some point of variability within a population. The human genome consists of many variations between individuals, as many as 150 million[4], but this only comprises 1% of the genome[5]. Each distinct variation in a marker is referred to as an *allele*, and the marker is said to be *biallelic* if it has two possible states (and triallelic for three, quadrillelic for four, etc.)
- **Co-dominance** - If a marker is not polymorphic in a unique way, then no information can be inferred between individuals and all modes of inheritance would be equally likely. Co-dominance asserts that all possible states of a marker are distinct from one another.

4 As of dbSNP version 146
5 Echo 'grep -c -oP "[ACTGactg]{1}" chr*.fa | awk -F:'{print $2}' | sed 's/\n/\+/g'' | bc

- **Hardy-Weinberg Equilibrium** - The HWE is the model that assumes that allele and genotype frequencies remain constant in the absence of external interference. A biallelic marker with a minor allele frequency of *p* and a major allele frequency of *q = (1-p)* will distribute genotypes with a $p^2 : 2pq : q^2$ ratio.
- **Low mutation rate** – In order to be certain that the marker found in an individual was inherited from a parent, the marker must have a low rate of mutation in the general population.

1.3.1 Markers

It is always beneficial to genotype all individuals of the same family using the same type of marker to ensure consistency by using the same set of genotype loci that are crucial in following the disease locus across generations.

Molecular markers are based upon the type of variation within the DNA they are recording, and variants typically come in two sorts: tandem repeats, and point mutations.

1.3.1.1 Tandem Repeats

These are sections of the genome where one or more nucleotides are repeated in succession an unspecified amount of times[6]. The nature of these repeats is unknown, but they are thought to be a historical remnant of viruses inserting their genetic material into our ancestor's genomes in aeons past (ref). Nonetheless, they serve as useful markers since they are an inheritable and observable, and thus informative for linkage.

There are several classes of tandem repeats, each class determined by either/both the number of repeats and the length of the repeating sequence. A *minisatellite* or *variable number tandem repeat*

6 Though we should assume a minimum of at least three repeats for a pattern to be detected.

(VNTR) is any repeating sequence approximately 10-60 nucleotides in length, and a *microsatellite* or *short tandem repeat (STR)* is typically fewer than 10 nucleotides.

Tandem repeats are extremely polymorphic, leaving very little ambiguity in which parent an allele was inherited from. However, they tend to reside in non-coding regions of the genome, which makes their resolution multi-gene specific, and not exon-specific.

1.3.1.2 Point Mutations

A point mutation or *single nucleotide polymorphism (SNP)* denote a variation in the genome of a single base pair. Despite the potential for a SNP to be quadrilellic most are biallelic, lending a lower level of informativeness than repeat markers. Their advantage lies in the sheer density of their numbers, covering the genome at approximately 200bp intervals within coding and non-coding regions alike. With SNPs, a disease locus can be identified at the inter-gene level, and it is common to have two or more SNPs within the same exon of a gene.

<XXX:Robert - Should I mention indels here? Are they even relevant to haplotype reconstruction?>

The low level of polymorphism does indeed pose a problem when trying to trace which parent a genotype descended from, and is one of main points of contention in this thesis.

1.3.1.3 **dbSNP**

Classically the minor allele frequencies of SNPs was said to be no less than 1% to have useable information content, but the most current SNP database (dbSNPv142) also caters for SNPs with much lower frequencies due to typically large sample sizes used in the genotyping process. The general format for SNP IDs is the reference SNP (or 'rs') identifier followed by a unique number that distinguishes it from other SNPs (e.g. rs1234567). Due to the large number of SNPs being

registered into dbSNP from several different sources, it is not uncommon to see two different SNPs specifying the same variant, or for a SNP with such a low minor allele frequency (<0.0001%) that it should hardly be called a SNP at all. Indeed, dbSNP rigorously strives to resolve these problematic SNPs with every release, but this leads to inconsistencies between versions and binds the informative of any SNP study to a specific dbSNP release version.

1.3.2 Linkage Maps

Linkage maps hold subsets of markers thought to be informative for the particular study in question. There are two main types of maps - physical and genetic, each with their own uses but for ultimately different purposes.

1.3.2.1 Physical Maps

Physical maps ensure that markers represent actual points on the genome, e.g. rs1234567 → chr7:97795920. The variant can be found by selecting the correct chromosome and offsetting the number of the base pairs from the start of the chromosome (chromosome 7 in this case).

This is very helpful when wanting the know the actual position of a mutation, but is of no use in linkage studies.

1.3.2.2 Genetic Maps

Genetic maps work on a fundamentally different principle to physical maps. Where physical maps infer actual base-pair distances between adjacent markers, genetic maps use recombination frequencies to predict the occurrence of crossovers. Genetic maps use units of centiMorgan as

opposed to base-pairs, and though both scale linearly in accordance with one another, there are notable differences when dealing with genders due to the different recombination frequencies.

The recombination frequencies between markers in genetic maps are utilized effectively by the linkage analysis for determining statistically accurate occurrences of crossover events between from parent to offspring. This information along with a known inheritance model is paramount in precisely determining where points of recombination could occur.

1.4 Modes of Inheritance

Inheritance models dictate the pathways that alleles segregate across generations. An individual with the same allele across both homologous chromosomes is said to be *homozygous*, whereas an individual with differing alleles at these homologs are said to be *heterozygous*. To delve into how these types of alleles interact, we must first stop to appreciate the foundation of modern genetics that these modes were built upon: Mendelian inheritance.

1.4.1 Mendel's Laws

Gregor Mendel was a botanist monk who outlined the first basic principles of genetics through his breeding experiments with peas. Though some of his observations have not stood incontestable through the passages of time, they are still of great import and he is said to be the 'father of genetics':

1.4.1.1 Law of Segregation

Every diploid organism has two alleles that segregate during gamete formation to create diploid offspring that bears just *one* of the alleles from that parent (i.e. each offspring inherits one allele from their parents). As shown in Figure 1, this has the interesting effect such that two homozygous parents with genotypes *ww* and *RR*

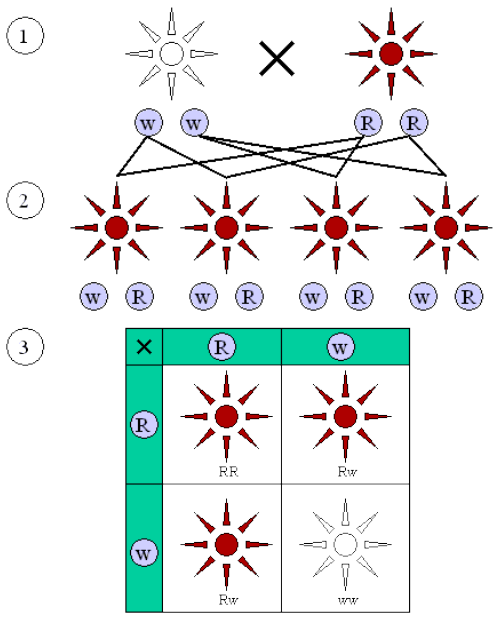

Figure 1: Image courtesy of Wikimedia Commons

respectively will *always* have heterozygous offspring with *wR* genotypes, yet these heterozygous offspring can further mate to reconstitute their parent's genotypes.

1.4.1.2 Law of Independent Assortment

Separate alleles allude to separate traits, and their transmission from parent to offspring are independent of one another. A zygote can end up with any combination of parental alleles, and may end up with genotypes that are entirely different from their parents. Offspring that (by chance) have the same genotypes as any of their parents are called *recombinant*, and offspring with differing genotypes from any of their parents are called *non-recombinant*.

1.4.1.3 Law of Dominance

If the presence of a single form of an allele is enough to produce the phenotype, then that heterozygous allele is said to be *dominant* (as is the case with *wR*, where the *R* trait dominates the *w* trait). Conversely, if a single form of an allele is not enough to produce the phenotype, then that heterozygous allele is said to be *recessive* (as is the case with *wR*, where the *w* trait is dominated by the *R* trait and cannot be expressed).

1.4.2 Mendel's Laws Revised

Though Mendel was ahead of his time in describing genes as allelic traits, he did not take into account that genes from diploid individuals may have more than two alleles (i.e. *multi-allelic*) due to variations within a population. He also made the erroneous assumption[7] that one allele maps to one trait, where in reality there are numerous *polygenic traits* produced by the interaction of many alleles.

In relation to his third law, the concept of *codominance* is missed; if the two potential phenotypes within heterozygous alleles are independent of each other and do not conflict, they can both be expressed. Alternately, if the two potential phenotypes within heterozygous alleles are capable of

7 One that likely every geneticist makes when first entering the field!

conflicting or interacting with one another, then the phenotype is usually a 'blend' of the two, and said to be of *incomplete dominance* since neither allele is completely dominant over the other.

1.4.1 Pedigrees

Individuals from the same family share a *pedigree*, and each individual is grouped into two classes: *founders* and *non-founders*.

Founders are individuals who contribute unique *founder alleles* to the pedigree, and the non-founders are the individuals who inherit them.

The larger the pedigree the more complex it is to process, with a function that is referred to as the *bit size (b)*.

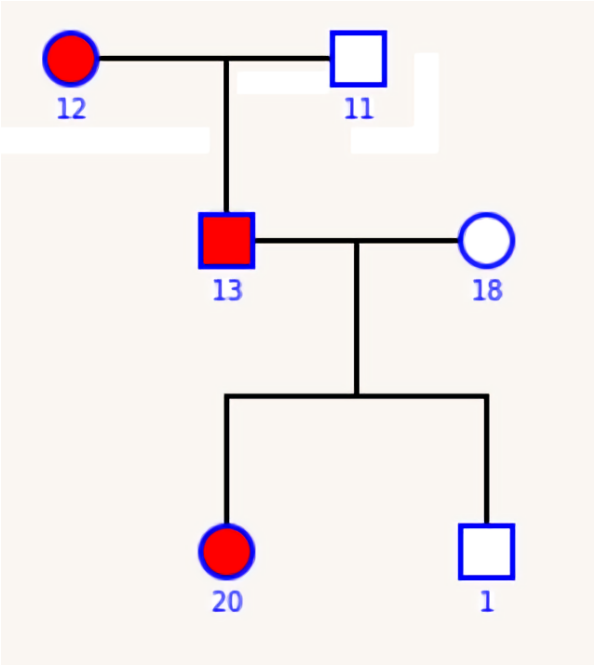

Figure 2: Typical pedigree, with females represented as circles and males as squares. Affectation is represented by a red colour fill.

$$b = 2n - f - g \qquad (2)$$

where:

n is the number of non-founders

f is the number of founders

g is the number of ungenotyped founder couples (zero if unknown)

Inbred pedigrees where related individuals have produced offspring are said to be *consanguineous.*

1.4.2 Inheritance Models

Inheriting a disease allele is not a straightforward process, since the disease may not be visible in every individual who has the allele. For a given individual it is assumed that the disease status is always known[8].

There is an empirical probability that they will present with the disease phenotype, given a set of disease alleles. This is the derived through what is known as a *penetrance function:*

$$P=(D|A)$$

where:

- D is disease status = affected or unaffected
- A is the genotype belonging to [GG, Gg, gg], and G is the disease allele.

1.4.2.1 Autosomal Dominant

If the disease segregates through an autosomal dominant model, then the disease trait is dominant and the penetrance function is thus: $P(D|GG) = P(D|Gg) = 1$, and $P(D|gg)=0$

Affected individuals are present at all generations, and because the disease trait lies within the autosomes (chromosomes 1 to 22), affectation is non-gender specific and both sexes have equal probability in inheriting the trait.

1.4.2.2 Autosomal Recessive

Autosomal recessive is a more constrained mode, where instead the disease trait is recessive and the penetrance function is limited to: $P(D|GG)=1$, and $P(D|Gg) = P(D|gg) = 0$

Here, affected individuals appear to skip generations since parents carrying a heterozygous genotype will not present the phenotype, but instead be *carriers* of the disease trait that can be reconstituted as a homozygous disease allele in their offspring. In consanguineous pedigrees, the likelihood of this happening is significantly increased due to the multiple pathways the disease allele can take, with each generation having potentially double the chance of inheriting the disease allele after each subsequent consanguineous pairing.

8 At least in binary traits which are present at birth.

1.4.2.3 X-linked Dominant

The disease trait rests solely on chromosome X. Due to males having only one X-chromosome and one Y-chromosome, this simplifies the inheritance pattern such that males can only inherit the single Y-chromosome from their father, and females can only inherit the single X-chromosome from their father – removing much ambiguity. The male's X-chromosome can still be one of the two maternal X-chromosomes, and likewise the female's maternal X-chromosome can be one of the two as well. X-linked dominant pedigrees can be identified by a distinct lack of father-to-son transmission, since that would imply that the Y-chromosome is the disease allele[9]. Due to dominant nature of the trait, the penetrance function is identical to that of autosomal dominant.

1.4.2.4 X-linked Recessive

In X-linked recessive diseases, both of the inherited X-chromosomes must contain the disease allele (i.e. be homozygous), with the exception of males who need only copy to express the phenotype since the X-chromosome inherently dominates the Y-chromosome. Males cannot be carriers in X-linked pedigrees making them *autozygous*.

Typically the pedigree exhibits more affected males than females, with skipped generations where the mother is only a carrier.

The penetrance function is gender-specific, taking on the same rules for females as autosomal recessive, but differing for males:

$$P(D|GG)=1 \text{ and } P(D|Gg) = P(D|gg)=0, \text{ if female}$$

$$P(D|GG)=N/A, \quad P(D|Gg)=1, \text{ and } P(D|gg)=0, \text{ if male}$$

9 <XXX: Robert – do I need to mention Y-inheritance? Does such a thing exist?>

1.5 Haploblocks

An individual's genotype is usually acquired in an *unphased* manner, meaning that that genotyping process acquired the alleles present at a locus, but the ordering was unimportant so that the maternal and paternal allele are ambiguous.

A genotype with a known path of inheritance for its alleles is said to be *phased*, and a phased genotype is alsc known as a *haplotype*. A haplotype typically spans a single locus similar to an allele, but a haplotype is encapsulated and a restricted to a larger group of haplotypes known as a *haploblock*, which represent the intervals between two recombination events[10].

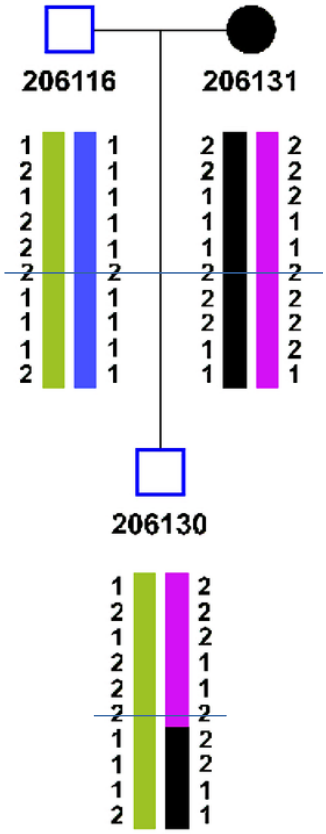

Figure 3: Four complete haploblocks from two founders inherited by their non-founder offspring. The sixth marker is highlighted due to ambiguous genotypes.

These blocks represent the splitting of founder alleles which occurs at every meiosis, effectively halving the founder data within the allele at each generation.

A founder allele stretching across *m* markers and undergoing *n* meiosis, will have an expected length of *m/n* markers over *n* generations.

In Figure 3 the offspring is recombinant at that particular locus due to a crossover in the maternal alleles. Note that at the level of a single marker it is not clear which way a genotype is phased, e.g. the sixth marker where the maternal alleles (2,2) and the paternal alleles (2,2) leave much ambiguity to which particular chromosome the offspring inherited their alleles (2,2).

10 Including the recombination that occurs at the start/end of each chromosome

Resolving ambiguous genotypes to specific chromosomes requires providing the genotype with some neighbouring context such that it knows whether it is in a recombination hotspot, or whether it lies within an existing block.

It is the aim of this thesis to derive methods to explore such problems, but in order to do so we must first understand how haplotypes are generated through linkage analysis.

2 Linkage Analysis and Haplotype Generation

In this chapter we look at the underlying principles that power linkage analysis; the general methodology, as well as the algorithms best suited for generating haplotypes from incomplete genotypes.

2.1 Theory of Linkage Analysis

In the previous chapter we discussed how genotypes are mapped to specific markers, and that a disease locus is usually only flanked by the nearest adjacent markers which may or may not be within the same haploblock, and thus may or may not be co-segragated by the disease allele during meiosis.

If a given marker co-segregates with a disease allele over multiple recombination events, then it is safe to assume that they are in close proximity of one another. Such bound loci are said to be *in linkage* with one another, and when one undergoes a meiosis; so does the other.

In linkage analysis, this translates as pinpointing the multiple subsets of markers with the fewest number of crossovers between themselves, such that the the number of meioses in the cases is minimized and the controls are maximized (i.e. which markers/traits are in linkage with one another in affected individuals, but are not in linkage within unaffected individuals).

Classical linkage analysis involves doing this by direct means; counting each crossover event for each marker against every other marker, but the complexity of such a task scales exponentially for large sets and is almost highly impractical.

Modern linkage analysis determines this using statistical means; through the filter of the correct genetic model to establish a correct penetrance function, the assumption of linkage equilibrium between markers to simplify computation by discounting crossover interference, the lack of monozygotic twins[11], and the known recombination frequencies between markers as well as their

11 Identical genotypes in twins bias the analysis since meioses may not be independent.

ordering. These are then used to establish a likelihood score that rates each marker upon the aforementioned principles such that disease loci can be found by scanning for the highest likelihood.

2.1.1 Likelihood and LODs

This likelihood is known as a *logarithm of the odds (LOD)* score, and is a ratio that compares the likelihood of any two traits being in linkage compared to them not being in linkage.

$$LOD(\theta)=\log_{10}\left(\frac{L(\tilde{\theta})}{L(\theta=0.5)}\right) \tag{3}$$

where:

- θ = ‰(1 – e $^{-2d}$), the Haldane map function that estimates the recombination frequency given the genetic distance in centiMorgan (as defined in equation (1) on page 7).
- θ = 0.5, the probability that two traits segregate independently (are essentially unlinked).
- $L(\theta)$ = the joint likelihood.

The joint likelihood is main core of what the linkage analysis tries to solve, iterating over all individuals and markers, as defined by:

$$L=\sum_{g1}\cdots\sum_{gn}\prod_{i}P(Y_i|g_i)\prod_{j}P(g_j)\prod_{\{k,l,m\}}P(g_m|g_k,g_l) \tag{4}$$

where:

- g_n = genotypes of individual n
- Y_n = phenotype of individual n
- $P(Y_i \mid g_i)$ = penetrance probability applied over all individuals. This is the probability of a phenotype given a genotype.

- $P(g_j)$ = founder probability applied to founders only. Modelled under Hardy-Weinberg and assumes linkage equilibrium.
- $P(g_m \mid g_k, g_l)$ = transmission probability. Probability of non-founder genotypes (m) given parental genotypes (k and l).

There is a narrow window of relevance when dealing with LOD scores, typically any score less than -2 suggests complete lack of linkage and though may seem uninformative, is actually a good indicator of where a disease locus *cannot* be.

A LOD score greater than or equal to 3 is considered to be of great significance for pinpointing a disease locus, but the reasoning for this is mostly historical.

2.2 Linkage Algorithms

Though the methodology between algorithms varies somewhat, all work under the same principle of determining a LOD score and generating missing genotypes, though some are more engineered towards resolving phase than others.

2.2.1 Elston-Stewart

The very first linkage algorithm first computed upon paper, by performing a task known as *peeling* upon a pedigree.

A connected graph of individuals (as nodes) and relationships (as edges) would be recursively peeled in a depth-first manner, by collapsing LOD score calculations onto members of a pedigree known as *pivots*, who are key to the pedigree such that their removal would create two detached pedigrees.

At each stage, the algorithm considers a simple nuclear family (mother, father, offspring(s)) and performs a LOD score computation of the offspring inheriting their founder parents genotypes, then flattening the score up a generation into that founder parent.

This eventually resorts in a single node that contains the total LOD summation of the entire pedigree for a given marker.

Over time, programs such as **Linkage** allowed multiple markers to be considered at the same time and gave rise to *multi-point* linkage analysis that became the standard for subsequent linkage packages.

The algorithm scales linearly with the number of individuals in a pedigree, but exponentially with the number of markers for a multi-point analysis; a severe limiting step in an age where the density of markers were on the increase.

Generating fast and reasonable accurate LOD results became the main aim for a lot of programs, each performing a different approximation upon the LOD score calculations to generate a faster analysis. The program **Vitesse** was boasted as being able to analyse 8 markers jointly using “fuzzy inheritance”.

Since haplotype reconstruction requires precise determination of a missing genotype, the Elston-Stewart algorithm was lacking in this regard.

2.2.2 Lander-Green

The era of haplotypes began with the Lander-Green algorithm, which aims to describe a probability distribution over all possible ways traits may segregate between individuals based on their genotypes for a given marker, whilst being dependent on adjacent markers.

Under Lander-Green, a pedigree is represented using an *inheritance vector*, which shows the

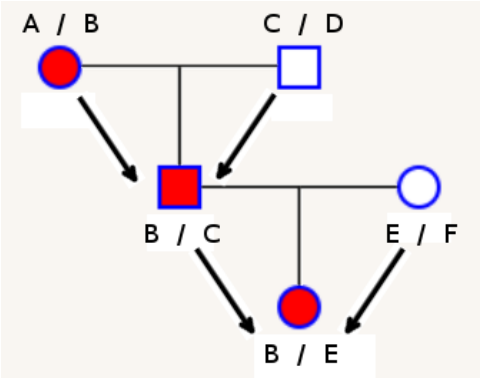

Figure 4: A pedigree of phased individuals with inheritance vectors represented as arrows. Here the order of the genotypes are {maternal,paternal} indexed to {0,1}

transition of founder alleles to non-founders.

The compact representation of the inheritance vector in Figure 4 would be 1000 since it is only the transition from the first generation mother→ child where the second allele (B→ 1) is transmitted, whereas it is always the first allele(0) in all other cases.

The importance of this for haplotype reconstruction is apparent, for it considers the complete inheritance probability distribution and any bisection of this set would capture a consistent set of genotypes as described by their vectors.

This was the first linkage analysis that made use of a Hidden Markov Model (HMM), which denotes a dependent chain of unobserved hidden states with state probabilities, that transition between each other under transmission probabilities.

Kruglyak readily applied this to a a linkage model where the states mapped the possible inheritance vectors, and the transmission probabilities were the genotype likelihoods. The Elston-Stewart algorithm was then used to sum over the inheritance probability distribution at each locus.

Over time, a few notable speedups were incorporated without the cost of accuracy incurred by approximation:

- **Inheritance Reduction** – where potential inheritance vectors with incompatible genotypes between parent and offspring can be pruned from the analysis.
- **Founder Symmetry** – where inheritance vectors that differ only in a founder's phase are equal, which lends massive importance in the regard that most pedigrees do not set phase at all (Allegro).
- **Founder Couple Reduction** – where ungenotyped founder couples are equivalent to each other and need only be evaluated once (Allegro).

Each of these cut the complexity of a pedigree in half, together producing a significant (~6x) reduction in bit-size.

The Lander-Green algorithm scales linearly with the number of markers, a much better time complexity than the Elston-Stewart, however suffers drawbacks in the number of individuals in the pedigree where it scales exponentially with each meiosis.

2.2.3 Hidden Markov Models and Haplotype Reconstruction

The use of connected graphs in linkage analysis has the side benefit of being able to reconstruct haplotypes within the same run.

In the HMMs, the linkage program transitions between a subset of inheritance vector states to determine the total likelihood by the time the last state in the chain is reached. The path taken by the linkage program is consistent; such that any paths that do not reach the last inheritance vector terminated early and were thus incompatible, but any paths that did make it through were conformed with the genotypes, and all the (now phased) genotypes that came before.

That is not to say that there is only one true path[12] in the HMM that is compatible with the observed genotypes; several may exist. One way of finding the optimal path is to run an inference algorithm through the HMM chain to find the most likely sequence of states given a likelihood.

2.2.3.1 The Viterbi Algorithm

One such algorithm is the forward-backward Viterbi algorithm. A forward-backward algorithm works on the principle of performing two passes upon a HMM. Given a sequential set of observations:

1. **Forward Pass** – Starts at the first state and steps forwards towards the end state, computing *forward probabilities* that set the probabilities of any given state ending at any other subsequent state. In linkage, this is essentially setting a weighted graph based on compatible inheritance vectors from the observed genotypes.
2. **Backwards Pass** - After the forward pass, computes a set of *backward probabilities* that sets the probabilities of any given state observing any applicable previous state. This is essentially the same as the previous step

12 At the molecular level there is only one true path, since the crossovers occur at discrete loci.

Once performed, a combined probability distribution is obtained to find the mostly likely state at any step in the chain.

The Viterbi algorithm is an extension of this prototype that attempts to maximize the probabilities such that the optimal path can be traced by observing the most likely set of inheritance vectors upon the most likely set of founder alleles.

This is still an approximation upon the data, since the most likely set of inheritance vectors only tells us how the alleles flow from generation to generation, but not necessarily the content of the alleles.

2.3 Summary

We have seen that there are two main limiting steps in linkage analysis algorithms that scale either with the number of meiosis or the number of markers. The Elston-Stewart algorithm's complexity scaled linearly with the number of meiosis by collapsing computations on pivot individuals, but scaled exponentially with the number of markers.

The Lander-Green(-Kruglyak) algorithm was the inverse of this; where the number of meiosis caused an exponential increase in complexity due to the possible inheritance vector states doubling at each step. The number of markers scaled linearly however, and this is what made it the most widely-adopted algorithm in linkage analysis programs, since the density of genotyping chips have only increased over the time.

Hidden Markov Models lend their use to inheritance or descent diagrams that allow us to set phase at each genotype by following the most likely set of inheritance vectors.

Haplotype reconstruction is by no means deterministic, as the optimal path as derived from the Viterbi algorithm may be just as likely as another path, especially for incomplete data where the potential alleles of untyped genotypes may have equal likelihood.

The strength of the analysis ultimately lies in the observed genotypes, and some are more informative than others, as we explore in greater detail within the next chapter.