Modelling genes: mathematical and statistical challenges in genomics

Peter Donnelly

Abstract. The completion of the human and other genome projects, and the ongoing development of high-throughput experimental methods for measuring genetic variation, have dramatically changed the scale of information available and the nature of the questions which can now be asked in modern biomedical genetics. Although there is a long history of mathematical modelling in genetics, these developments offer exciting new opportunities and challenges for the mathematical sciences. We focus here on the challenges within human population genetics, in which data document molecular genetic variation between different people. The explosion of data on human variation allows us to study aspects of the underlying evolutionary processes and the molecular mechanisms behind them; the patterns of genetic variation in different geographical regions and the ancestral histories of human populations; and the genetic basis of common human diseases. In each case, sophisticated mathematical, statistical, and computational tools are needed to unravel much of the information in the data, with many of the best methods combining complex stochastic modelling and modern computationally-intensive statistical methods. But the rewards are great: key pieces of scientific knowledge simply would not have been available by other means.

Mathematics Subject Classification (2000). Primary 92D10, 92D15; Secondary 65C05.

Keywords. Genetics, problems related to evolution, Monte Carlo methods.

1. Introduction

We begin with a brief review of the basic concepts and terminology from genetics. The full picture is both more complicated and richer than we need, and we present only a very high-level overview.

Genetic information is transmitted from parents to offspring, and carried in the nucleus of each cell, in DNA (deoxyribonucleic acid). To a mathematician, DNA can be thought of as a very long word over the four letter alphabet $\{A, C, G, T\}$, with each letter representing one of the four chemical bases, or nucleotides, which are arranged effectively linearly along the DNA molecule. It is the order in which the bases appear which conveys the information. Some parts of the molecule are "read" by molecular machinery, and the relevant part of the DNA is used as a template to make a particular protein. These parts of the DNA are called genes. The totality of an organism's DNA is called its genome. The human genome consists of about 3×10^9 bases, and contains around 25,000 genes, but most of the DNA in the human genome appears to have no

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

function. The genomes of different organisms differ in size, with some much smaller and some substantially larger than the human genome. Like many other organisms, humans are diploid, in that we carry two copies of our genome, one inherited from our mother, and one from our father. Human DNA is packaged into 23 pairs of chromosomes, with one copy of each chromosome inherited from each parent.

Each human sperm or egg (collectively referred to as germ cells) contains a single copy of each of the 23 chromosomes. For what follows, we need to understand a little about the process during which germ cells are formed, called meiosis. Focus on a particular human chromosome. The individual (progenitor) producing the germ cell will have two (slightly different) copies of this chromosome. Think of the process which produces the chromosome for the germ cell as starting on one of the chromosomes in the progenitor and copying from it along the chromosome. Occasionally, and for our purposes randomly, the copying process will "cross over" to the other chromosome in the progenitor, and then copy from that, perhaps later jumping back and copying from the original chromosome, and so on. The chromosome in the germ cell will thus be made up as a mosaic of the two chromosomes in the progenitor. The crossings over are referred to as recombination events. In a typical human meiosis, there will be only a few recombination events per chromosome. In addition to the process of recombination, there will be very occasional mutations: positions where the nucleotide in the offspring is different from that in the progenitor chromosome from which it is being copied. To give an idea of the scale of these effects in humans, the probability of a mutation in any particular nucleotide position is of order 10^{-8} per meiosis, and the average probability of a recombination event in a particular position is of the same order. Mutation and recombination are two of the fundamental evolutionary forces. Mutation introduces new variants into a population. (Some of these will make the resulting chromosome better at doing its job than the progenitor chromosome). The effect of recombination is more subtle, but equally important. Recombination allows the shuffling of variants between different backgrounds: when a mutation arises it occurs on a particular chromosome with a particular DNA sequence. Over generations, recombination events near this mutation allow it to be swapped onto different backgrounds.

In effect, the human genome project read one copy of the human genome ([12], [26]) – actually a mosaic made up of the genome from many individuals. The human genome sequence is available on the web, along with annotations which show, for example, which parts correspond to known and predicted genes, or regions which regulate the expression of genes, or appear to be highly conserved across species. The genomes of many other organisms are also now available, with more being completed each month. In each case a major challenge within the science is to better understand the function of, and interactions between, different parts of each of these genomes.

We can think of the human genome project as focussing on the aspects of our genome which we all share: the things that make us human. But there are also differences between people, in appearance, nature, abilities, and susceptibility to different diseases. Some of these differences have a genetic component, resulting from differences in the DNA sequence between individuals. If we compared two human chromosomes in the same region then they would differ at about 1 place in 1000. (As a comparison, the human genome sequence differs from that of the chimpanzee at about 1 position in 100.)

Following on from the human genome project, there was a major effort in a public-private partnership to discover many of the positions at which human chromosomes differ. While there are a number of interesting ways in which DNA sequences can differ, the most common is when at a particular position, or nucleotide, some chromosomes in the population carry one letter (or base) while others carry a different letter (base). Such positions are called single nucleotide polymorphisms, or SNPs (pronounced "snips"). These SNPs are catalogued in public databases (e.g. http://www.bioinfo.org.cn/relative/dbSNP%20Home%20Page.htm). Over the current decade the number of SNPs known in humans has grown from hundreds to more than 8 million. Nonetheless, many remain undiscovered. For example it is estimated that there are 10 million "common" SNPs, that is SNPs where the rarer variant has a population frequency of at least 5% [11]. As noted above, the mutation rate at any particular nucleotide position is very small (10^{-8}). On the other hand the human genome is large (3×10^9 nucleotides). SNPs can be thought of as positions at which mutations happen to have occurred in the genome, where the chromosome carrying the mutation has spread through the population. It is extremely rare for there to be more than two variants present at a particular SNP.

Having read one copy of the human genome sequence, then catalogued many of the DNA sequence variations present in human populations, a natural next step was to understand the patterns in which these variants occur in different populations. This has recently been undertaken by the International HapMap Consortium, a collaboration involving five different countries (Canada, China, Japan, UK, USA), at a cost of about $100M. Largely completed, the project typed around 3.5M SNPs in samples from four populations around the world: 90 Caucasians from Utah, 90 Yorubans from Ibadan, Nigeria, 45 Han Chinese from Beijing, and 45 Japanese from Tokyo. The first phase of the project, involving just over 1M SNPs, was reported in [11].

It turns out that the variants present at SNPs close to each other on the same chromosome are often correlated. That is, if at one SNP some chromosomes in the population carry an A and others a G, while at a nearby SNP the two variants are T and C, it might be that chromosomes which carry an A at the first SNP are more likely to carry a T at the second SNP than those with a G at the first SNP. This kind of correlation is known in population genetics as linkage disequilibrium (LD). The correlations can be very strong (for example the extreme case where all chromosomes carry either A and T or G and C, is not unusual) and as we will see below, are very important for studies of the genetics of human disease. Amongst other things, the HapMap project characterized patterns of linkage disequilibrium in the samples it studied.

The reasons for linkage disequilibrium are apparent when one thinks about the history of novel mutations. In the example above, suppose the mutation giving rise to the first SNP (A/G) occurred further into the past than that giving rise to the second SNP, and suppose that at the second SNP the C variant (variants are often called alleles in genetics) was the one present originally. The mutation creating a T and giving rise to the second SNP will have occurred on a single chromosome in a particular generation. Suppose it occurred on a chromosome carrying an A at the first SNP. Then it is immediate that when it arose, a T at the second SNP would occur with an A at the first SNP. Over subsequent generations, the number of copies of the chromosome carrying the T is likely to have grown (otherwise it would not be present today) and unless there is a recombination event between the two SNP positions on one of these chromosomes, it will remain the case that a T at the second SNP would always occur with an A at the first SNP. This association will only be broken down by recombination events, and the extent of this will depend on two things: (i) how many nucleotides separate the two SNPs on the chromosome (the closer together, on average, the smaller is the chance of a recombination between them); and (ii) the number of generations since the mutation giving rise to the second SNP (since a larger number of generations will allow a greater chance for a recombination event). In general, the observed patterns of LD depend on a number of factors, including chance past recombination events, and the demographic history of the population concerned.

2. Mathematical models

The arguments in the previous section were entirely qualitative. While helpful, they do not allow quantitative assessments of the way in which various aspects of genetic variation depend on the underlying evolutionary forces or demographic effects. To do so requires the development and analysis of mathematical models of the evolutionary process.

There has been a long history of mathematical modelling in population genetics, dating back to Fisher and Wright early last century. For most questions of interest, stochastic effects are important and the principal models are probabilistic. For most of the period over which these models have been studied, empirical data against which to compare the models have been sparse. Typically, what data there were came from so-called model organisms (particular species of flies and worms for example). Over the last few years, there has been an explosion of data documenting genetic variation in humans, to the extent that our own species provides the richest setting in which to apply these models.

We aim here only to give a brief flavour of the stochastic models which arise in population genetics. The most basic models are finite Markov chains which describe the way in which the genetic composition of the population changes over time. In most cases, these models are not tractable, and interest moves to their limiting behaviour as the population size grows large, under suitable re-scalings of time. When examined forward in time, this leads to a nice family of measure-valued diffusions, called

Fleming–Viot processes. In a complementary, and for many purposes more powerful approach, one can instead look backwards in time, and focus on the genealogical tree relating sampled chromosomes. In the large population limit, these (random) trees converge to a particular process called the coalescent.

One simple discrete model for population demography is the Wright–Fisher model. Consider a population of fixed size N chromosomes which evolves in discrete generations. (For many purposes it turns out that we can ignore the fact that chromosomes occur in pairs in individuals, and we do so here.) The random mechanism for forming the next generation is as follows: each chromosome in the next generation chooses a chromosome in the current generation (uniformly at random) and copies it, with the choices made by different chromosomes being independent. An equivalent description is that each chromosome in the current generation gives rise to a random number of copies in the next generation, with the joint distribution of these "offspring numbers" being symmetric multinomial. Under an assumption of genetic neutrality, all variants in a population are equally fit. In this case, one can first generate the demography of the population using, say, the Wright–Fisher model, and then independently superimpose the genetic type for each chromosome, and the details of the (stochastic) mutation process which may change types. The extent to which this neutrality assumption applies is rather controversial in general, and for humans in particular, but it seems likely that it provides a reasonable description for many parts of the genome. Recombination (and if needed natural selection) can be naturally added to the model. Where a recombination event occurs, the offspring chromosome will be made up from two chromosomes in the current population. Although we have described it in terms of chromosomes, it is natural only to apply the Wright–Fisher model to small regions of a chromosome. In this case, the probabilities of mutation and recombination in a copying event are both very small, and these events are rare.

The Wright–Fisher model may also be extended to allow for more realistic demographic effects, including variation in population size, and geographical spatial structure in the population (so that offspring chromosomes are more likely to be located near to their parents). We will not describe these here. Somewhat surprisingly, it transpires that the simple model described above, (constant population size, random mating, and neutrality – the so-called "standard neutral" model), or rather its large population limit, captures many of the important features of human evolution. There is an aphorism in statistics that "all models are false, but some are useful". The standard neutral model has proved to be extremely useful.

In a Wright–Fisher or any other model, we could describe the genetic composition of the population at any point in time by giving a list of the genetic types currently present, and the proportion of the population currently of each type. Such a description corresponds to giving a probability measure on the set E of possible types. It is sometimes helpful to think of this measure as the distribution of the type of an individual picked at random from the population. In this framework, when we add details of the mutation process and recombination to the Wright–Fisher model, we obtain a discrete time (probability) measure-valued Markov process. As N becomes large a suitable rescaling of the process converges to a diffusion limit: time is measured in units of N generations, and mutation and recombination probabilities are scaled as N^{-1}. For general genetic systems, the limit is naturally formulated as a measure-valued process, called the Fleming–Viot diffusion. The classical so-called Wright–Fisher diffusion is a one dimensional diffusion on [0, 1] which arises when there are only two genetic types and one tracks the population frequency of one of the types. This is a special case of the Fleming–Viot diffusion, in which we can identify the value of the classical diffusion, $p \in [0, 1]$ with a probability measure on a set with just two elements. The beauty of the more general, measure-valued, formulation is that it allows much more complicated genetic types, which could track DNA sequences, or more exotically even keep track of the time since particular mutations arose in the population.

The Fleming–Viot process can thus be thought of as an approximation to a large population evolving according to the Wright–Fisher model. For the Wright–Fisher model, time is measured in units of N generations in this approximation (and the approximation applies when mutation and recombination probabilities are of order N^{-1}). In fact the Fleming–Viot process arises as the limit of a wide range of demographic models, (and we refer to such models as being within the domain of attraction of the Fleming–Viot process) although the appropriate time scaling can differ between models. (See, for example, [5].) For background, including explicit formulations of the claims made above, see for example [2], [3] [4], [5], [6]. Donnelly and Kurtz ([2], [3]) give a discrete construction of the Fleming–Viot process. As a consequence, the process can actually be thought of as describing the evolution of a hypothetically infinite population, and it explicitly includes the demography of that population.

There has been considerable recent interest in looking backwards in time to study the genealogy of population genetics models. This is simplest in the absence of recombination. Consider again the discrete Wright–Fisher model. If we consider two different chromosomes in the current generation, they will share an ancestor in the previous generation with probability $1/N$. If not, they retain distinct ancestries, and will share an ancestor in the previous generation with probability $1/N$. The number of generations until they share an ancestor is thus geometrically distributed with success probability $1/N$ and mean N. In the limit for large N, with time measured in units of N generations, this geometric random variable will converge to an exponential random variable with mean 1.

More generally, if we consider k chromosomes, then for fixed k and large N, they will descend from k distinct ancestors in the previous generation with probability

$$1 - \binom{k}{2}\frac{1}{N} + O(N^{-2}).$$

Exactly two will share a common ancestor in the previous generation with probability $\binom{k}{2}\frac{1}{N} + O(N^{-2})$ and more than a single pair will share a common ancestor with probability $O(N^{-2})$. In the limit as $N \to \infty$, with time measured in units of N generations, the time until any of the k share an ancestor will be exponentially distributed with mean $\binom{k}{2}^{-1}$, after which time a randomly chosen pair of chromosomes will share an ancestor.

Thus, in the large population limit, with time measured in units of N generations, the genealogical history of a sample of size n, may be described by a random binary tree. The tree initially has n branches, for a period of time T_n, after which a pair of branches (chosen uniformly at random independently of all other events) will join, or coalesce. More generally, the times T_k, $k = n, n-1, \dots, 2$ for which the tree has k branches are independent exponential random variables with

$$\mathrm{E}(T_k) = \binom{k}{2}^{-1},$$

after which a pair of branches (chosen uniformly at random independently of all other events) will join, or coalesce. The resulting random tree is called the n-coalescent, or often just the coalescent.

In a natural sense the tree describes the important part of the genealogical history of the sample, in terms of their genetic composition. It captures their shared ancestry, due to the demographic process. A key observation is that in neutral models the distribution of this ancestry is independent of the genetic types which happen to be carried by the individuals in the population. Probabilistically, one can thus sample the coalescent tree and then superimpose genetic types: first choose a type for the most recent common ancestor of the population (the type at the root of the coalescent tree) according to the stationary distribution of the mutation process, and then track types forward through the tree from the common ancestor, where they will possibly be changed by mutation.

The preceding recipe gives a simple means of simulating the genetic types of a sample of size n from the population. Note that this is an early example of what has recently come to be termed “exact simulation”: a finite amount of simulation producing a sample with the exact distribution given by the stationary distribution of a Markov process. In addition, it is much more computationally efficient than simulating the entire population forward in time for a long period and then taking a sample from it. Finally, it reveals the complex structure of the distribution of genetics models at stationarity – the types of each of the sampled chromosomes are (positively) correlated, exactly because of their shared ancestral history.

We motivated the coalescent from the Wright–Fisher model, but the same limiting genealogical tree arises for any of the large class of demographic models in the domain of attraction of the Fleming–Viot diffusion. Moreover, the way in which the tree shape changes under different demographic scenarios (e.g. changes in population size, geographical population structure) is well understood.

The discrete construction of the Fleming–Viot process described above actually embeds the coalescent and the forward diffusion in the same framework, so that one can think of the coalescent as describing the genealogy of a sample from the diffusion.

There is even a natural limit, as $n \to \infty$ of the n-coalescents. This can be thought of as the limit of the genealogy of the whole population, or as the genealogy of the infinite population described by the Fleming–Viot process, although the analysis underlying the relevant limiting results is much more technical than that outlined above for the fixed-sample-size case. It is easiest to describe this tree from the root, representing the common ancestor of the population, forward to the tips, each of which represents an individual alive at the reference time. The tree has k branches for a random period of time T_k, after which a branch, chosen uniformly at random, independently for each k, splits to form two branches. The times T_k, $k = 2, 3, \dots,$ are independent exponential random variables, and independent of the topology of the tree, with

$$\mathrm{E}(T_k) = \binom{k}{2}^{-1}.$$

Write

$$T = \sum_{k=1}^{\infty} T_k$$

for the total depth of the tree, or equivalently for the time back until the population first has a common ancestor. Note that T is a.s. finite. In fact $\mathrm{E}(T) = 2$.

Now we return to the case where recombination is allowed. The simplest way to conceptualise this more general situation is that there is a genealogical tree, marginally distributed as the coalescent, associated with each nucleotide position. As one moves along the DNA sequence, these trees for different positions are highly positively correlated. In fact, two neighbouring positions will have the same tree iff there is no recombination event between those positions since their joint most recent common ancestor, on a lineage leading to the current sample. If there is such a recombination, the trees for the two positions will be identical back to that point, but (in general) different before it. The correlation structure between the trees for different positions is complex, and for example regarded as a process on trees as one moves along the sequence, it is not Markov. But it is straightforward to simulate from the relevant joint distribution of trees, and hence of sampled sequences. The trees for each position can be embedded in a more general probabilistic object (this time a graph rather than a tree) called the ancestral recombination graph ([8], [9]).

3. Disease mapping

One major current analytical challenge in the field is the development of statistical methods in genetic studies of human disease. A common study design is case-control: a (typically large) set of individuals with a particular disease (cases) and a set of healthy individuals (controls) are typed at a (large) set of SNPs. If one variant at a particular SNP predisposes individuals to (respectively protects them against) the disease in question then the frequency of that variant should be higher (lower) in the cases than the controls. The signal one looks for then is a difference in allele frequency between cases and controls at a particular SNP.

As one contemporary example, the Wellcome Trust Case Control Consortium is a large UK-based study in which 2000 cases for each of 8 common diseases (Type 1 and Type 2 Diabetes, Hypertension, Coronary Heart Disease, Crohn's Disease, Bipolar Disorder, Rheumatoid Arthritis, and Tuberculosis) will be compared with 3000 controls at around 500,000 SNPs. This size of study is becoming more common, and although expensive, is within reach of major biomedical research budgets. (For example, the study just described will cost around US$15M.)

Some human disorders have a simple genetic component. In these, a single gene will be involved, and mutations in that gene cause individuals to be affected. In some cases, such as Huntington's disease, individuals will be affected if either of their chromosomes carries the mutation. (The inheritance is said to be dominant.) In others, such as Cystic Fibrosis, individuals will be affected only if both their chromosomes carry mutations at the gene in question. (The inheritance is said to be recessive.) In these cases there is effectively a deterministic relationship between carrying mutated copies of the gene in question and having the disease. These so-called simple genetic diseases are typically rare, and often very debilitating. In a large number of cases the exact genes involved are now known.

Most or all of the common human diseases also have a genetic component, but one which acts in a more subtle, and complicated, way. We are some way from understanding the full story, but for these common human diseases, it is thought that mutations in genes may slightly increase (or decrease) the probability of the individual developing the disease, rather than deterministically predicting it. Disease susceptibility may well also involve the interaction between mutations in different genes, and/or interactions between genes and environmental or lifestyle factors.

One major issue with case-control studies involves the need for cases and controls to be as similar as possible apart from their disease status. A particular, genetic, concern relates to geographical population structure. Most human populations differ genetically – individuals from nearby geographical locations are more likely to be genetically similar than those from distant locations. This is well documented in comparisons between the major continental regions of the world. But the same effect pertains, to a lesser extent, within continental or even national regions. Suppose for simplicity that a population is actually made up of two subpopulations which differ genetically at a particular genetic marker (say the A allele is more common in subpopulation 1 than in subpopulation 2), and that in addition, perhaps for environmental reasons, the disease is more common in subpopulation 1. Then a random sample of cases from the population as a whole will tend to include more individuals from subpopulation 1 than will a random sample of controls, and in turn the sample of cases will have a higher frequency of A than will the controls. A naive analysis, which ignores the population substructure, might wrongly conclude that the A variant at this SNP played a role in disease susceptibility.

This tendency for geographical population structure to lead to false positives in association studies actually led to the case-control design being largely ignored for many years. More recently a range of statistical approaches has been developed to correct the problem. One class of approach uses all the markers typed to correct the null distribution of the usual test statistics. Another uses the markers to infer the underlying structure and assign individuals to subpopulations, with the comparisons between cases and controls being made only within subpopulations. Perhaps counterintuitively, it is also the case that the problems caused by substructure increase with the size of the study: even the small amounts of structure within national populations might cause problems for the large studies currently being undertaken. See [19] for further discussion and additional references.

If an association study directly tests a SNP causatively involved in disease susceptibility, then we would expect to see frequency differences between cases and controls at that SNP. For common diseases the effect of carrying one variant is typically small, which will lead to only a small difference in the frequency of that variant between cases and controls. The large sample sizes of current studies are needed to ensure statistical power to detect small frequency differences.

Even were an investigator to restrict attention only to variants which occur at appreciable frequency (e.g. so-called "common variants", where the less common, or "minor" allele has frequency $> 5\%$), these cannot all be tested in an association study. It is estimated that there are probably around 10 million such variants in the human genome [11]. Firstly, many of these variants are not known, and secondly, the cost of checking all known variants is prohibitive (even at levels of current biomedical research funding).

Here, the correlations between alleles described above (recall the discussion of linkage disequilibrium) is very helpful. In an extreme case, suppose variants at two SNPs are perfectly correlated: each chromosome carrying an A at the first SNP carries a T at the second, and chromosomes with a G at the first SNP carry a C at the second. In this case, when an association study types one of the SNPs it effectively also types the other. Put another way, if one of the two SNPs were causatively involved in the disease, and a study typed the other SNP, then there should still be a signal.

In fact, this "extreme" case, of perfect correlation, is not uncommon. For example, it is estimated that in a Caucasian population, 60% of common SNPs have the property that there are at least three other SNPs with which they are perfectly correlated, and 20% are perfectly correlated with more than 20 other SNPs; only around 20% are not perfectly correlated with any other SNPs [11]. (There is in general less correlation between SNPs in samples from African populations.) The reasons for this are now well understood. They follow from properties of coalescent trees, and recently discovered facts about the human recombination process. Two SNPs will be perfectly correlated iff they occur on the same branch of the coalescent tree. The branches near the root of the tree are relatively long, thus allowing time for a number of mutations to occur. In addition, as we discuss in more detail below, it turns out that in humans, recombination events do not occur uniformly along the chromosome, but instead cluster into small regions, called recombination hotspots, which are typically widely separated. Between these regions there will often be effectively a single coalescent tree for all nucleotide positions, and the placement of mutations on branches of this tree induces the correlations between SNPs.

Often two SNPs may be well correlated but not perfectly so. In this case, if one is causative and the other is typed, it may still be possible to see a signal, and hence to detect the untyped causative SNP. Under a simple disease model, this effect depends simply on the correlation coefficient, r^2, between the SNPs. If a sample size of n were needed at given power to detect the SNP in a study in which it is typed directly, then a sample size of n/r^2 will be needed for the same power if only the correlated SNP were typed.

The major point of the HapMap project was to describe these correlations between SNPs in human populations. As a consequence, association studies can carefully choose which SNPs to type so as to minimize the number of SNPs involved. For example, it is estimated that in a Caucasian sample, genotyping a set of 300,000 SNPs will capture around 80% of all common SNPs with $r^2 > 0.8$ [11].

Whatever strategy is chosen to select SNPs for typing in an association study, the analytical challenge is how best to analyse the data. This can helpfully be thought of as a statistical missing data problem. We have data at a set of SNPs in cases and controls. For these SNPs we can directly test the possibility that their variants are related to disease susceptibility in a variety of ways (e.g. by simple chi-squared tests for differences in genotype frequencies, or by fitting say logistic regression models relating genotype to disease status). If we had data at the SNPs not typed in the study we could apply the same tests at those SNPs. Thinking of the data at the untyped SNPs as missing data, the challenge then is to learn about the missing data from the data we actually have. Some of the most promising approaches to this problem make use of the mathematical models described above, (or approximations to them). Informally, the models can be used to predict, or impute, data at the untyped SNPs from the data at the SNPs actually typed, and then this imputed data is used to test for a disease effect. The approach can be applied either at the positions of known SNPs not included in the study, or more generally at arbitrary positions in the genome. It combines the empirical information available from surveys such as HapMap, inferences as to recombination rates in the human genome, and population genetics models.

4. Human recombination

Recall that recombination is the process by which germ cells are constructed to contain part of each of the chromosomes in their progenitor. It has long been known that recombination events do not happen uniformly along the human chromosomes. (The rates of recombination even differ between males and females.) Over large scales, this can be seen in pedigree (or family) studies: effectively through localising the genomic positions of recombination events in comparisons between parents and their children.

But pedigree studies have limited resolution for estimating recombination rates. The average recombination rate across 10 million basepairs – 10 megabases (Mb) –

is 10^{-1}. Reliable estimation of probabilities of order 10^{-1} requires many tens, or hundreds of observations. While this is realistic in human pedigree studies, the average recombination rate across 1Mb is an order of magnitude lower, and requires an order of magnitude more observations for accurate estimation, taking it effectively beyond the limit of practicability for pedigree studies. As a consequence, our understanding of the variation in human recombination rates based on pedigree studies does not go below the megabase scale.

So what do recombination rates look like over finer scales. Two recent lines of evidence suggested that the picture may be surprising and very interesting. The first was the direct observation of recombination hotspots: small (typically 2 kilobase, or kb) regions in which recombination events cluster, and for which the local recombination rate is much higher than in the surrounding DNA. These observations typically involved studies of human sperm. Although realistic pedigree studies are uninformative over these scales, clever and careful experimentation does allow detection of sperm with recombination events in particular small regions, from which recombination rates (in males) can be estimated over the region studied.

The second clue came from patterns of linkage disequilibrium (LD) in human populations. Contrary to the predictions of simple models, human linkage disequilibrium extended over much larger regions than expected, and in addition, the patterns showed regions of extended LD interrupted by short regions of LD breakdown. Although both the initial observations and the suggested causes were controversial, this pattern is now well documented in humans [11]. One natural explanation was that recombination events were not uniformly distributed but instead clustered into hotspots.

As we saw above, the patterns of genetic variation in human populations have been shaped by a number of effects, including recombination. In principle then, such data contain information about the underlying recombination rates. Armed with an understanding of the stochastic models of section 2, we could then treat this as a statistical problem, and try to use data to infer some of the parameters of the models, in particular the recombination rates.

It turns out that this is a challenging inference problem, for a number of reasons. In either classical or Bayesian statistical inference, a central role is played by the likelihood: the probability of the observed data as a function of model parameters. Although the stochastic models are well understood, and for example easy to simulate from, no explicit expressions are available for probabilities of interest, such as the stationary distribution. In the statistical context, this means that the likelihood associated with the model is not available analytically. One way of conceptualising the difficulty is as follows: for a given genealogical tree (or graph in the context of recombination) one could calculate the likelihood for given parameter values. The actual likelihood could then be obtained by averaging this quantity over all possible underlying trees or graphs. Herein lies the problem: the space of trees/graphs is so large that this averaging is impracticable.

Various clever computational approaches have been developed in modern statistics to overcome this type of problem, and there has been particular attention to these in the population genetics context (see for example [25]). A general observation is that sophisticated understanding of the stochastic models allows big improvements in the quality and efficiency of the statistical inference. For example, for inference of recombination rates assumed constant across the region of interest, the best available methods are more efficient then their predecessors by up to four orders of magnitude [7].

But even the best available statistical methods based on the coalescent are impracticable for estimating recombination rates of interest for a different reason: the sheer size of available data sets. The Phase I HapMap data, for example, documents genetic variation at 1M SNPs in 269 individuals. Two different approaches have recently been developed to address this issue. In essence, each takes the view that the coalescent model is itself an approximation to reality, so why not make further judicious approximations in order to achieve tractability. One approach, pioneered by Li and Stephens [18] involves an alternative model for genetic data with a hidden Markov structure. (See [1] for application to the estimation of fine-scale recombination rates.) We concentrate here on the other approach, introduced by Hudson [10] in the context of constant recombination rates, and developed by McVean and colleagues for variable recombination rates [20], [21].

This approach retains the original coalescent model, and uses its exact likelihood for data at a pair of SNPs, as a function of the recombination rate between them. But rather than using the correct joint distribution for a set of more than two SNPs, the approximation assumes each pair of SNPs to be independent. In this way, a so-called pairwise composite likelihood is constructed as the product of the exact coalescent likelihoods across all pairs of SNPs in the data. In the setting of variable recombination rate, the parameter of interest is a piecewise constant function specifying the recombination rate between each pair of SNPs (so the function only changes value at the positions of SNPs). McVean *et al.* [21] adopt a Bayesian approach to inference. The prior distribution (here on function space) encourages smoothness in the rate function and reversible jump Markov chain Monte Carlo, using the pairwise composite likelihood, is used to explore the posterior distribution on recombination rates. One attractive feature of the prior distribution putting weight on smooth functions is that the approach "borrows" information from nearby SNPs – the estimated rate between a pair of SNPs will be influenced by data at nearby SNPs. Another positive consequence is the tendency to avoid overfitting: maximum likelihood estimation with the same likelihood function would fit a different recombination rate between each pair of SNPs. The paper also develops a formal likelihood ratio test for the presence of a recombination hotspot, based on the composite likelihood (with significance levels determined by appropriate simulation). Tests on real and simulated data show these methods to perform remarkably well. In spite of the approximations (firstly to give the coalescent and secondly of the coalescent) involved, the models appear to be capturing key features of the real world remarkably well.

These new statistical methods, applied to recent genome-wide variation data sets such as the HapMap, have enormously extended our knowledge of human recombination. For the first time, fine-scale estimates of recombination rates are available across the human genome. These so-called genetic maps have resolution 2-3 orders of magnitude finer than their predecessors from pedigree studies. They show striking variation in rates, by up to four orders of magnitude, over kilobase scales, and provide a powerful tool in studies of human disease. We now know that recombination hotspots are a ubiquitous feature of the human genome: whereas around 15-20 hotspots had been previously characterised, around 30,000 have been detected from the new approach, with an estimated average density of one hotspot per 50kb. The approach has demonstrated, also for the first time, that recombination hotspots are definitively a feature of female recombination, and more generally that the fine-scale recombination landscape appears similar between males and females. Contrary to previous reports, recombination rates are systematically lower within genes, but interestingly, systematically higher close to genes [11], [22].

One common general tool in genetics studies involves comparisons between species, and this has proved informative for recombination as well. Two comparisons between humans and our closest neighbouring species, the chimpanzee, revealed that recombination hotspots are a feature of chimpanzee recombination as well. But whereas human and chimpanzee DNA sequences agree at about 99% of nucleotide positions, the studies surprisingly found that the positions of recombination hotspots do not match between the two species, suggesting that they have evolved rapidly over evolutionary times [29], [23]. Comparisons between the time-averaged recombination rates estimated from population data and those in contemporary sperm suggest an even more rapid evolution of recombination hotspots, with substantial changes even over the half a million or so years over which human genetic variation has accumulated [13].

Perhaps the best example of mathematical and statistical approaches adding substantially to scientific knowledge in this area comes from studies of motifs associated with recombination hotspots. The question of why some parts of the DNA sequence act as recombination hotspots and some do not has been a major focus of research attention, and remains little understood. No clear pattern was available from the 15 hotspots directly characterised from human sperm typing. But by first identifying, and then studying, 25,000 hotspots, Myers *et al.* [22] were able to identify several short DNA sequence motifs associated with hotspots. (In fact the analysis was not straightforward, essentially because most hotspots are localised only to within 5-10kb by statistical methods. The key was to focus only on those hotspots containing specific sequences of several hundred basepairs – many such so-called repetitive elements abound in the human genome – and to compare them with the same sequences outside hotspots.) Although also an exciting ongoing story, this approach was the first to identify sequence motifs (one important example is the collection $CCTCCCT$ of eight basepairs) associated with human hotspots, and amongst the first to be identified for any organism.

5. Conclusion

In addition to explaining some of the science, our aim has been to give a sense of the central role being played in modern genomics by mathematical modelling and statistical methods. The Human Genome Project provided the foundation for a new generation of genetics research. As we build on that foundation, in our understanding of basic biology, of the genetic basis for disease susceptibility, and in the use of this information to develop new therapies and preventions for human disease, it is clear that the mathematical and computational sciences will continue to play a vital role.

References

[1] Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., Stephens, M., Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics* **36** (2004), 700–706.

[2] Donnelly, P., Kurtz, T. G., A countable representation of the Fleming-Viot measure-valued diffusion. *Ann. Probab.* **24** (2) (1996), 698–742.

[3] Donnelly, P., Kurtz, T. G., Particle representations for measure-valued population models. *Ann. Probab.* **27** (1) (1999), 166–205.

[4] Ethier, S. N., Kurtz, T. G., *Markov processes. Characterization and convergence.* Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., John Wiley, New York 1986.

[5] Ethier, S. N., Kurtz, T. G., Fleming-Viot processes in population genetics. *SIAM J. Control Optim.* **31** (2) (1993), 345–386.

[6] Ewens, W. J., *Mathematical population genetics. I. Theoretical introduction*. Second edition, Interdiscip. Appl. Math. 27, Springer-Verlag, New York 2004.

[7] Fearnhead, P., Donnelly, P., Estimating recombination rates from population genetic data. *Genetics* **159** (2001), 1299–1318.

[8] Griffiths, R. C. and Marjoram, P., Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3** (1996), 479–502.

[9] Griffiths, R. C. and Marjoram, P., An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (ed. by Peter Donnelly and Simon Tavaré), IMA Vol. Math. Appl. 87, Springer-Verlag, Berlin 1997, 257–270.

[10] Hudson, R. R., Two-locus sampling distributions and their application. *Genetics* **159** (2001), 1805–1817.

[11] The International HapMap Consortium, A haplotype map of the human genome. *Nature* **437** (2005), 1299–1320.

[12] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409** (2001), 860–921.

[13] Jeffreys A. J., Neumann R., Panayi M., Myers S., Donnelly P., Human recombination hot spots hidden within regions of strong marker association. *Nature Genetics* **37** (2005), 601–606.

[14] Kingman, J. F. C., Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics* (Rome, 1981), North-Holland, Amsterdam, New York 1982, 97–112.

[15] Kingman, J. F. C., The coalescent. *Stochastic Process. Appl.* **13** (3) (1982), 235–248.

[16] Kingman, J. F. C., On the genealogy of large populations. Essays in statistical science. *J. Appl. Probab.* Special Vol. **19A** (1982), 27–43.

[17] Kurtz, T. G., Martingale problems for conditional distributions of Markov processes. *Electron. J. Probab.* **3** (9) (1998), 29 pp. (electronic).

[18] Li, N., Stephens, M., Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* **165** (2003), 2213–2233.

[19] Marchini, J., Cardon, L., Phillips, M., Donnelly P., The effects of human population structure on large genetic association studies. *Nature Genetics* **36** (2004), 512–517.

[20] McVean, G. A. T., Awadalla, P., Fearnhead, P., A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160** (2002), 1231–1241.

[21] McVean, G. A. T., Myers, S., Hunt, S., Deloukas, P., Bentley, D. R., Donnelly, P., The fine-scale structure of recombination rate variation in the human genome. *Science* **304** (2004), 581–584.

[22] Myers S., Bottolo L., Freeman C., McVean G., Donnelly P., A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310** (2005), 321–324.

[23] Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A., Pääbo, S., Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics* **37** (2005), 429–434.

[24] Saunders, I. W., Tavaré, S., Watterson, G. A., On the genealogy of nested subsamples from a haploid population. *Adv. in Appl. Probab.* **16** (3) (1984), 471–491.

[25] Stephens, M., Donnelly, P., Inference in Molecular Population Genetics. *J. Royal Statist. Soc. Ser. B* **62** (2000), 605–655.

[26] Venter, J. C. *et al.*, The sequence of the human genome. *Science* **291** (2001), 1304–1354.

[27] Watterson, G. A., Mutant substitutions at linked nucleotide sites. *Adv. in Appl. Probab.* **14** (2) (1982), 206–224.

[28] Watterson, G. A., Substitution times for mutant nucleotides. Essays in statistical science. *J. Appl. Probab.* Special Vol. **19A** (1982), 59–70.

[29] Winckler W., Myers S. R., Richter D. J., Onofrio R. C., McDonald G. J., Bontrop R. E., McVean G. A. T., Gabriel S. B., Reich D., Donnelly P., Altshuler D., Fine-scale recombination rates differ markedly in human and chimpanzee. *Science* **308** (2005), 107–111.

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
E-mail: donnelly@stats.ox.ac.uk