

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray
Copyright © 2003 John Wiley & Sons, Ltd.
ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

SECTION 1

**AN INTRODUCTION TO
BIOINFORMATICS FOR THE
GENETICIST**

CHAPTER 1

Introduction: The Role of Genetic Bioinformatics

MICHAEL R. BARNES¹ and IAN C. GRAY²

¹*Genetic Bioinformatics and* ²*Discovery Genetics
Genetics Research Division
GlaxoSmithKline Pharmaceuticals, Harlow, Essex, UK*

- 1.1 Introduction
 - 1.2 Genetics in the post-genome era — the role of bioinformatics
 - 1.3 Knowledge management and expansion
 - 1.4 Data management and mining
 - 1.5 Genetic study designs
 - 1.5.1 The linkage approach
 - 1.5.2 The association approach
 - 1.5.3 Markers for association studies
 - 1.6 Physical locus analysis
 - 1.7 Selecting candidate genes for analysis
 - 1.8 Progressing from candidate gene to disease-susceptibility gene
 - 1.9 Comparative genetics and genomics
 - 1.10 Conclusions
 - References
-

1.1 INTRODUCTION

In February 2000, scientists announced the draft completion of the human genome. If media reports were accepted at face value, then it might be reasonable to predict that most geneticists would be unemployed within a decade of this announcement and human disease would become a distant memory. As we all know this is very far from the truth, the human genome is many things but it is not in itself a panacea for all human ailments, nor is it a revelation akin to the elucidation of the DNA double helix or the theory of evolution. The human genome is simply a resource borne out of technical prowess, perhaps with a little human inspiration. One thing that is certain is that we do not yet understand the functional significance of the majority of our genome, but what we do know is finally put into context. Over the past 200 years mankind has developed an

ever increasing understanding of genetics; Darwin and Mendel provided the 19th century theories of evolution and inheritance, while Bateson, Morgan and others established a framework for the mechanisms of genetics at the beginning of the 20th century. The tentative identification of DNA as the genetic material by Avery and colleagues in the 1940s preceded the elucidation of the structure of the DNA molecule in 1953 by Watson and Crick, which in turn provided a mechanism for DNA replication and ushered in the era of modern molecular genetics. In 2003, precisely 50 years after this landmark discovery it is anticipated that the entire human genome sequence will be completed in a final, polished form; a fully indexed but currently only semi-intelligible 'book of life'. Here lies the most overlooked property of the genome—its value as a framework for data integration, a central index for biology and genetics. Almost any form of biological data can be mapped to a genomic region based on the genes or regulatory elements that mediate it. So the sequencing of the human genome means new order for biology. This order is perhaps comparable to the order the periodic table brought to chemistry in the 19th century. Where elements were placed in an ordered chemical landscape, biological elements will be grouped and ordered on the new landscape of the human genome. This presents excellent opportunities to draw together very diverse biological data; only then will the 'book of life' begin to make sense.

The human genome and peripheral data associated with and generated as a result of it require increasingly sophisticated data storage, retrieval and handling systems. With the promises and challenges that lie ahead, bioinformatics can no longer be the exclusive realm of the Unix guru or the Perl hacker and in recent years web browsers have made tools accessible and user friendly to the average biologist or geneticist. Bioinformatics is now both custodian and gatekeeper of the new genome data and with it most other biological data. This makes bioinformatics expertise a prerequisite for the effective geneticist. This expertise is no mystery; modern bioinformatics tools coupled with an inquiring mind and a willingness to experiment (key requirements for any scientist, bioinformatician or not) can yield confidence and competence in bioinformatic data handling in a very short space of time. The objective of this book is not to act as an exhaustive guide to bioinformatics, other texts are available to fulfil this role, but instead is intended as a specialist guide to help the typical geneticist navigate the internet jungle to some of the best tools and databases for the job, that is, associating genes, polymorphisms and mutations with diseases and genetic traits. In this chapter we give a flavour of the many processes in modern genetics where bioinformatics has a major impact and refer to subsequent chapters for greater detail.

At the risk of over simplifying a very complex issue, the process of understanding genetic disease typically proceeds through three stages. First, recognition of the disease state or syndrome including an assessment of its hereditary character; second, discovery and mapping of the related polymorphism(s) or mutation(s) and third, elucidation of the biochemical/biophysical mechanism leading to the disease phenotype. Each of these stages proceeds with a variable degree of laboratory investigation and bioinformatics. Both activities are complementary, bioinformatics without laboratory work is a sterile activity as much as laboratory work without bioinformatics can be a futile and inefficient one. In fact these two sciences are really one, genetics and genomics generate data and computational systems allow efficient storage, access and analysis of the data—together, they constitute bioinformatics. Almost every laboratory process has a complementary bioinformatics process, Table 1.1 lists a few of these—building on these basic applications will maximize the effect of bioinformatics on workflow efficiency.

TABLE 1.1 Examples of Bioinformatics Applications in Genetics Research

| Data | Related Laboratory Techniques | Associated Bioinformatics Applications |
|-------------------------------------|--|--|
| Human genome sequence | DNA sequencing PCR Novel gene identification by expression analysis <i>In vitro</i> characterization of regulatory elements | Gene and regulatory region prediction BLAST homology searching Electronic PCR PCR primer design Electronic translation and protein secondary structure prediction |
| Genetic markers | Genotyping | <i>In silico</i> design of expression constructs Identification of optimal marker sets Genotyping assay design QC checking and statistical analysis of genotype data |
| Model organism genome sequence | Comparative genetics (e.g. linkage) and genomics (e.g. transgenics and gene knock-outs) | Linkage analysis of models of human diseases Comparative genetic and physical maps for cross-species analysis of linkage regions Functional assessment of gene regulatory regions by cross-species comparison <i>In silico</i> drafting of gene knock-out and transgenic constructs |
| Expression RNA and protein | Microarrays Serial analysis of gene expression (SAGE) Proteomics | Gene regulatory analysis Tumour and other disease tissue classification Elucidation of gene-gene interactions and disease pathway expansion |
| Three-dimensional protein structure | Crystallography/NMR | Prediction and visualization of molecular structures related to disease and mutation |

1.2 GENETICS IN THE POST-GENOME ERA – THE ROLE OF BIOINFORMATICS

In the role of genome data custodian and gatekeeper, bioinformatics is an integral part of almost every field of biology, including of course, genetics. In the broadest sense it covers the following main aspects of biological research:

- Knowledge management and expansion
- Data management and mining
- Study design and support
- Data analysis
- Determination of function

These categories are quite generic and could apply to any field of biology, but are clearly applicable to genetics. Both genetics and bioinformatics are essentially concerned with asking the right questions, generating and testing hypotheses and organizing and interpreting large amounts of data to detect biological patterns.

1.3 KNOWLEDGE MANAGEMENT AND EXPANSION

Few areas of biological research call for a broader background in biology than the modern approach to genetics. This background is tested to the extreme in the selection of candidate genes to test for involvement with a disease process, where genes need to be chosen and prioritized based on many criteria. Often biological links may be very subtle, for example a candidate gene may regulate a gene which regulates a gene that in turn may act upon the target disease pathway. Faced with the complexity of relationships between genes, geneticists need to be able to expand pathways and identify complex cross talk between pathways. As this process can extend almost interminably to a point where virtually every gene is a candidate for every disease, knowledge management is important to help to weigh up evidence to prioritize genes. The geneticist may not be an authority in the disease area under study, and in today's climate of reductionist biology an expert with a global picture of the disease process at the molecular level may be hard to find. Therefore effective tools are needed to quickly evaluate the role of each candidate and its related pathways with respect to the target phenotype.

Literature is the most powerful resource to support this process, but it is also the most complex and confounding data source to search. To expedite this process, some databases have been constructed which attempt to encapsulate the available literature, e.g. On-line Mendelian Inheritance in Man (OMIM). These centralized data resources can often be very helpful for gaining a quick overview of an unfamiliar pathway or gene, but inevitably one needs to re-enter the literature to build up a fuller picture and to answer the questions that are most relevant to the target phenotype or gene. The internet is also an excellent resource to help in this process. In Chapter 2, we offer some pointers to help the reader with effective literature searching strategies and give suggestions as to some of the best disease databases and related resources on the internet.

1.4 DATA MANAGEMENT AND MINING

Efficient application of knowledge relies on well organized data and genetics is highly dependent upon good data, often in very large volumes. Accessing available data, particularly in large volumes is often the biggest informatic frustration for geneticists. Here

we focus on aspects of accessing data from public databases; solutions for in-house data collection, either in the form of ‘off the shelf’ or custom-built laboratory information management systems (LIMS) belong to a specialist area that lies beyond the scope of this book.

Genetic data have grown exponentially over the last few years, fuelled by the expressed sequence tag (EST) cDNA sequence resources generated largely during the 1990s and more recently the increasing genomic sequence data from the human genome and other genome sequencing projects. Genetic database evolution has matched this growth in some areas, with some resources leading the efforts towards whole genome integration of genetic data, particularly the combined human genome sequence, genetic map, EST and SNP databases exemplified by the Golden Path. Curiously, development in many of the older more established genetic resources (for example, GDB and HGMD) has been somewhat stagnant. This may be partly due to the difficulties involved in data integration with the draft genome sequence, which is effectively a moving target as the data are updated on a regular basis. Many of the traditional genetic databases have not seized the opportunity to integrate genetic data with the human genome sequence. The future survival of these databases will certainly depend on this taking place and there is no question that the role of these databases will change. One might question the value of some of the older genetic datasets, for example, why would we need radiation hybrid maps of the human genome, when we have the ultimate physical map — the human genome sequence? These painstakingly collected datasets have already played a critical role in the process of generating the maps that allowed the sequencing of the human genome and they may still have some value as an aid for QC of new data and perhaps more importantly as a point of reference for all the studies that have previously taken place.

A key problem that frequently hinders effective genetic data mining is the localization of data in many independent databases rather than a few centralized repositories. A clear exception to this is SNP data which has now coalesced around a single central database — dbSNP at NCBI (Sherry *et al.*, 2001). By contrast human mutation data, which has been collected over many years, is still stored in disparate sources, although moves are afoot to move to a similar central database — Hobbies (Fredman *et al.*, 2002). These developments are timely; human mutation and polymorphism data both hold complementary keys to a better understanding of how genes function and malfunction in disease. The availability of a complete human genome presents us with an ideal framework to integrate both sets of data, as our understanding of the mechanisms of complex disease increase, the full genomic context of variation will become increasingly significant.

With the exception of dbSNP most recent database development has not been implicitly designed for geneticists, instead genomic databases and genome viewers have developed to aid the annotation of the human genome. Of course this data is vital for genetics, but this explains why the available tools often appear to lack important functionality. One has to make use of what functionality is available, although sometimes this means using tools in ways that were not originally intended (for example many geneticists use BLAST to identify sequence primer homology in the human genome, but few realize that the default parameters of this tool are entirely unsuited for this task). We will attempt to address these issues throughout this book and offer practical solutions for obtaining the most value from existing tools wherever possible. In Chapter 5 we examine the use of human genome browsers for genetic research. Tools such as Ensembl and the UCSC human genome browser annotate important genetic information on the human genome, including SNPs, some microsatellites and of course, genes and regulatory regions. User-defined queries place genes and genetic variants in their full genomic context, giving very

detailed information on nearby genes, promoters or regions conserved between species, including mouse and fish. It is difficult to overstate the value of this information for genetics. For example, cross-species genome comparison is invaluable for the analysis of function, as inter-species sequence conservation is generally thought to be restricted to a functionally important gene or regulatory regions and so this is one of the most powerful tools for identifying potential regulatory elements or undetected genes (Aparicio *et al.*, 1995). Several chapters in this book cover tools and databases to support these approaches (see Chapters 12 and 13).

As technology developments have scaled up the throughput of genotyping to enable studies of tens (and possibly hundreds) of thousands of polymorphisms and provided the capability to generate equally impressive amounts of microarray transcript data to name just two examples, the need for more effective data management has intensified. This reveals the major drawback of the ultra user-friendly ‘point and click’ interfaces to most genetics and genomics tools—they often do not allow retrieval of bulk datasets; instead data often has to be retrieved on a point by point basis. For many applications this is highly inefficient at best or simply non-viable at worst. One solution to this problem is to query the database directly at a UNIX or SQL level, but this may not be a trivial process for the occasional user with no or limited knowledge of command lines and in many cases it will not be possible to access the data directly in this manner. If the raw data are available, it may be possible to build custom databases, using database tools such as Microsoft ACCESS. However, the authors accept that this is not a straightforward option nor the method of choice of most users and instead this book will focus on web-based methods for data access. Where there is no web-based method to achieve a data mining goal, geneticists should consider contacting the developers of databases to request new functionality, such requests are generally welcomed by database developers, many of whom would be very pleased to know that their tools are being used! Several developers have already improved their methods for bulk data retrieval (probably as a result of requests from users), but interfaces are still lacking in some critical areas for genetics. For example, several tools allow the user to generate a list of SNPs across a locus (e.g. dbSNP, Ensembl and UCSC), but only one allows the user to retrieve the flanking sequence of each SNP in one batch to allow primer design (SNPper—see Chapter 3). We will try to tackle these problems as they arise throughout the book.

1.5 GENETIC STUDY DESIGNS

There are a number of approaches to disease gene hunting and many arguments to support the merits of one approach over another. Whatever the method, comprehensive informatics input at the study design stage can contribute greatly to the quality, efficiency and speed of the study. It can help to define a locus clearly in terms of the genes and markers that it contains and supports a logical and systematic approach to marker and gene selection and subsequent genetic analysis, simultaneously reducing the cost of a project and improving the chances of successfully discovering a phenotype–genotype correlation.

Despite the recent improvements in the throughput of genetic and genomic techniques and the increased availability of gene and marker data, genes which contribute to the most common human diseases are still very elusive. By contrast, the identification of genes mutated in relatively rare single gene disorders (so-called Mendelian or monogenic disorders) is now straightforward if suitable kindreds are available. The identification of the genes responsible for a plethora of monogenic disorders is one of the genetics

success stories of the late 1980s and the 1990s; genes identified include, to name but a few—*CFTR* (cystic fibrosis; Riordan *et al.*, 1989), Huntington (Huntington's disease; Huntington's Disease Collaborative Research Group, 1993), Frataxin (Friedreich's ataxia; Campuzano *et al.*, 1996) and *BRCA1* in breast and ovarian cancer (Miki *et al.*, 1994).

Unfortunately, success in the identification of genes with a role in complex (i.e. multi-genic) disease has been far less successful. Notable examples are the involvement of *APOE* in late-onset Alzheimer's disease and cardiovascular disease and the role of *NOD2* in Crohn's disease (Hugot *et al.*, 2001; Saunders *et al.*, 1993). However, genes for most of the common complex diseases remain elusive. Our ability to detect disease genes is often dependent on the analysis method applied. Methods for the identification of disease genes can be divided neatly into two broad categories, linkage and association. Although many common principles apply to both of these study types, each approach has distinct informatics demands.

1.5.1 The Linkage Approach

The vast majority of Mendelian disease genes have been identified by linkage analysis. This involves identifying a correlation between the inheritance pattern of the phenotypic trait (usually a disease state) with that of a genetic marker, or a series of adjacent markers. Because of the relatively low number of recombination events observed in the 2–5 generation families typically used for linkage analyses (around one per Morgan, which is roughly equivalent to 100 megabases, per meiosis), these marker/disease correlations extend over many megabases (Mb), allowing adequate coverage of the entire human genome with a linkage scan of only 300–600 simple tandem repeat (STR) markers giving an average spacing of 10 or 5 cM respectively. STRs are the markers of choice for linkage analysis, due to the fact that they show a high degree of heterozygosity. Markers with a heterozygosity level of >70% are typically selected for linkage panels (i.e. from 100 individuals selected at random, at least 70 would have two different alleles for a given marker; clearly the higher the heterozygosity the greater the chance of following the inheritance pattern from parent to offspring). Such marker panels are well characterized and can be accessed from several public sources at various densities (see Chapter 7). Just over 16,000 STR markers have been characterized in humans, which represents a small fraction of the estimated total numbers of polymorphic STRs. Analysis of the December 2001 human genome draft sequence suggests that there may be somewhere in the order of 200,000 potentially polymorphic STRs in the human genome (Viknaraja *et al.*, unpublished data). Software tools are now available to assist in the sequence-based identification of these potentially polymorphic STR markers across a given locus, should additional markers be required to narrow a linkage region (see Chapter 9 for details).

Clearly the limited degree of recombination that facilitates linkage analysis with sparse marker panels is a double-edged sword; the investigator may be left with several megabases of DNA containing a large number of potential candidate genes. However, combining data from several different families often results in reduction of the genetic interval under study, and the high-throughput sequencing capabilities available in many modern genetics laboratories coupled with complete genome sequence render the systematic screening of a large number of candidate genes a far less daunting task than it was 10 years ago.

Unlike single gene Mendelian diseases, complex genetic diseases are caused by the combined effect of multiple polymorphisms in a number of genes, often coupled with environmental factors. The successes of linkage analysis in the rapid identification of

Mendelian disease genes has spawned large-scale efforts to track down genes involved in the more common complex disease phenotypes. This approach is not restricted to academic research groups; many pharmaceutical and biotechnology companies have joined what many would perceive to be a 'genetic gold-rush', in an attempt to identify new drug targets for common diseases such as asthma, diabetes and schizophrenia, in a manner reminiscent of the rush to mine drug targets from expressed sequence tags (ESTs) in the late 1990s (Debouck and Metcalf, 2000). The application of a linkage approach to complex disease typically involves combining data from a large number of affected sib-pairs. Publicly available software for linkage analysis of sib-pairs is described in detail in Chapter 11.

Unfortunately the identification of genes involved in common diseases using a linkage strategy has been largely unsuccessful to date, mainly because each gene with phenotypic relevance is thought to make a relatively small contribution to disease susceptibility. These small effects are likely to be below the threshold of detection by linkage analysis in the absence of unfeasibly large sample sizes (Risch, 2000). In an attempt to circumvent this problem researchers using linkage approaches to identify genes involved in complex disease typically relax the threshold of acceptable 'log of the odds' (LOD) score (see Chapter 11) from 3, the traditionally accepted threshold of evidence for linkage in monogenic disease to 2, or sometimes even lower (Pericak-Vance *et al.*, 1998). However we would expect to see a number of hits due to chance alone with a comprehensive genome scan at this threshold. The rationale for lowering the threshold for detection of linkage, i.e. the effect of each contributing gene in a complex disease is smaller than would be expected for a monogenic disease, can result in a situation where a true signal is indistinguishable from background noise. In order to distinguish true linkage from false positives, many investigators are now using a combination of both linkage and association, relying on linkage analysis to reveal tentative, broad map positions which are subsequently confirmed and narrowed with an association study (see Chapter 8).

1.5.2 The Association Approach

In its simplest form, the aim of a genetic association study is to compare an allele frequency in a disease population with that in a matched control population. A significant difference may be indicative that the locus under test is in some way related to the disease phenotype. This association could be direct, i.e. the polymorphism being tested may have functional consequences that have a direct bearing on the disease state. Alternatively, the relationship between a genetic marker and phenotype may be indirect, reflecting proximity of the marker under test to a polymorphism predisposing to disease. The phenomenon of co-occurrence of alleles (in this case a disease-conferring allele and a surrogate marker allele) more often than would be expected by chance is termed linkage disequilibrium (LD). Suitable population structures for genetic association studies and statistical methods and software tools for the analysis of data resulting from such studies are discussed in detail in Chapters 8 and 11. Our aim here is to give the reader the briefest of introductions.

Association studies have three main advantages over linkage studies for the analysis of complex disease: (i) case-control cohorts are generally easier to collect than extended pedigrees; (ii) association studies have greater power to detect small genetic effects than linkage studies; a clear example is the insulin gene, which shows extremely strong association with type 2 diabetes, but very weak linkage (Spielman *et al.*, 1993); (iii) LD typically stretches over tens of kilobases rather than several megabases (Reich *et al.*, 2001), allowing focus on much smaller and more manageable loci. Among other reasons (discussed in

Chapter 8), this is because an association-based approach exploits recombination in the context of the entire population, rather than within the local confines of a family structure.

Of course, this last point is the other side of the double-edged sword of marker density and resolution mentioned in the context of linkage analysis above. The trade-off is reduced range over which each marker can detect an effect, resulting in a need for increased marker density. The required marker density for an association-based genome scan is unknown at present as we do not have enough information regarding human genome diversity in terms of polymorphic variability and genome-wide patterns of LD. However, typical guesses are in the range of 30,000–300,000 markers (Collins *et al.*, 1999; Kruglyak, 1999); orders of magnitude higher than the numbers required for linkage analysis. The high cost of generating the several million genotypes for such an experiment has prevented any such undertaking at the time of writing, although several proof of concept studies have demonstrated that high-density SNP maps can be efficiently generated using existing technologies and should be achievable in a reasonable time-frame (Antonellis *et al.*, 2002; Lai *et al.*, 1998). In the meantime, it is likely that research groups will continue to test individual genes for association with disease (the ‘candidate gene’ approach — see Section 1.7 below).

Once the genomic landscape, in terms of polymorphism and LD, is known with some degree of accuracy, it is highly likely that the number of markers required for a whole genome association study can be reduced by an intelligent study design with heavy reliance on bioinformatics input. Testing all available markers in a given region for association with a disease is expensive, laborious and frequently unnecessary; a simple example to illustrate this would be two adjacent markers which always demonstrate co-segregation; in other words, the genotypic status of one can always be predicted by genotyping the other—there is no point in genotyping both. Although this example is simple in the extreme, as adjacent markers typically show varying degrees of (rather than absolute) co-segregation, there is a trade-off between minimizing the amount of required genotyping whilst minimizing loss of information. Selection of optimal non-redundant marker sets, coupled with an initial focus on gene-rich regions, is the key to providing lower overall genotyping costs whilst retaining high power to detect association. This will require extensive knowledge of the blocks of preserved marker patterns (haplotypes) in the population under study; bioinformatics tools for constructing and analysing haplotypes and selecting optimal marker sets based on haplotypic information are discussed in detail in Chapters 8 and 11.

1.5.3 Markers for Association Studies

STRs were (and still are) the vanguard of linkage analysis, mainly because of their high levels of heterozygosity and hence increased informativeness when compared to an earlier marker system, the restriction fragment length polymorphism (RFLP); the majority of RFLPs are the result of a single nucleotide polymorphism (SNP) which creates or destroys a restriction site. SNPs have made a comeback worthy of Lazarus in recent years and are now the marker of choice for genetic association studies. The main reasons for the return to favour of SNPs are their abundance (an estimated 7 million with a minor allele frequency of greater than 5% in the human genome; Kruglyak and Nickerson, 2001) and binary nature which renders them well suited to automated, high-throughput genotyping. As mentioned above, tens or hundreds of thousands of SNPs will be required for whole genome association scans (even with optimized marker sets). Until very recently, studies on this scale were unfeasible, not only as a result of unacceptably high genotyping costs,

but also due to the lack of available markers. Large-scale SNP discovery projects such as the SNP consortium (TSC; Altshuler *et al.*, 2000a) have increased the number of known SNPs dramatically. We now have a great deal of SNP data (3.4 million non-redundant SNPs deposited in dbSNP at the time of going to press), however it is becoming apparent that even this number of markers will be insufficient for comprehensive association studies (note that the figure of 3.4 million includes a considerable number of SNPs with a minor allele frequency of less than 5%, which may be of limited use in association studies; this is discussed in Chapter 8).

We have already touched on the importance and potential impact of defining haplotypes as the basis for identifying optimal marker sets. This method has already been applied in small-scale studies with striking results. For example, in a study of nine genes spanning a total of 135 kb, Johnson *et al.* (2001) found that just 34 SNPs from a total of 122 could be used to define all common haplotypes (those with a frequency of greater than 5%) across the nine genes, an impressive validation of the approach of defining maximally informative minimal marker sets based on haplotypic data. However this study also highlighted the inadequacy of the current public SNP resource; only 10% of the SNPs identified by Johnson *et al.* were found to be present in dbSNP. Using dbSNP data alone, it was impossible to capture comprehensive haplotype data; in fact for four of the nine genes, no SNPs whatsoever were registered in dbSNP. Unfortunately it appears that our current public SNP resource represents the tip of the iceberg in terms of requisite information for the proper implementation of modest candidate gene association studies, let alone whole genome scans. However, given the burgeoning nature of dbSNP, we are optimistic that this situation is transient.

As a footnote to this section, it should be noted that although STRs have been largely swept aside by the wave of SNP euphoria, STRs may still be useful for association studies; indeed, it is possible that LD can be detected over far greater distances with STRs than SNPs under some circumstances, as discussed in Chapter 8.

1.6 PHYSICAL LOCUS ANALYSIS

In recent years, as the human genome sequence has neared completion, practical approaches to physical characterization of a genetic locus have changed quite dramatically. The laborious laboratory-based process of contig construction using yeast and bacterial artificial chromosome (YAC and BAC) clones or cosmids, involving consecutive rounds of library screening, clone characterization and identifying overlaps between clones, has become largely redundant, as has clone screening for the identification of novel polymorphic markers and genes. Today this process, which took many months or even years, can be completed in an afternoon using web-based human genome browsers. This shifts the initial focus of a study from contig construction and characterization to very detailed locus characterization using a range of bioinformatics tools; it is now possible to characterize a locus *in silico* to a very high level of detail before any further laboratory work commences. When the wet work does start, good prior use of bioinformatics will have rendered many procedures superfluous and the study is far more efficient and focused as a result. Figure 1.1 illustrates some of the key stages in the genetic analysis of candidate genes and loci—the role of informatics at each stage of this process is explored in detail in this book and the relevant chapters addressing each issue are indicated.

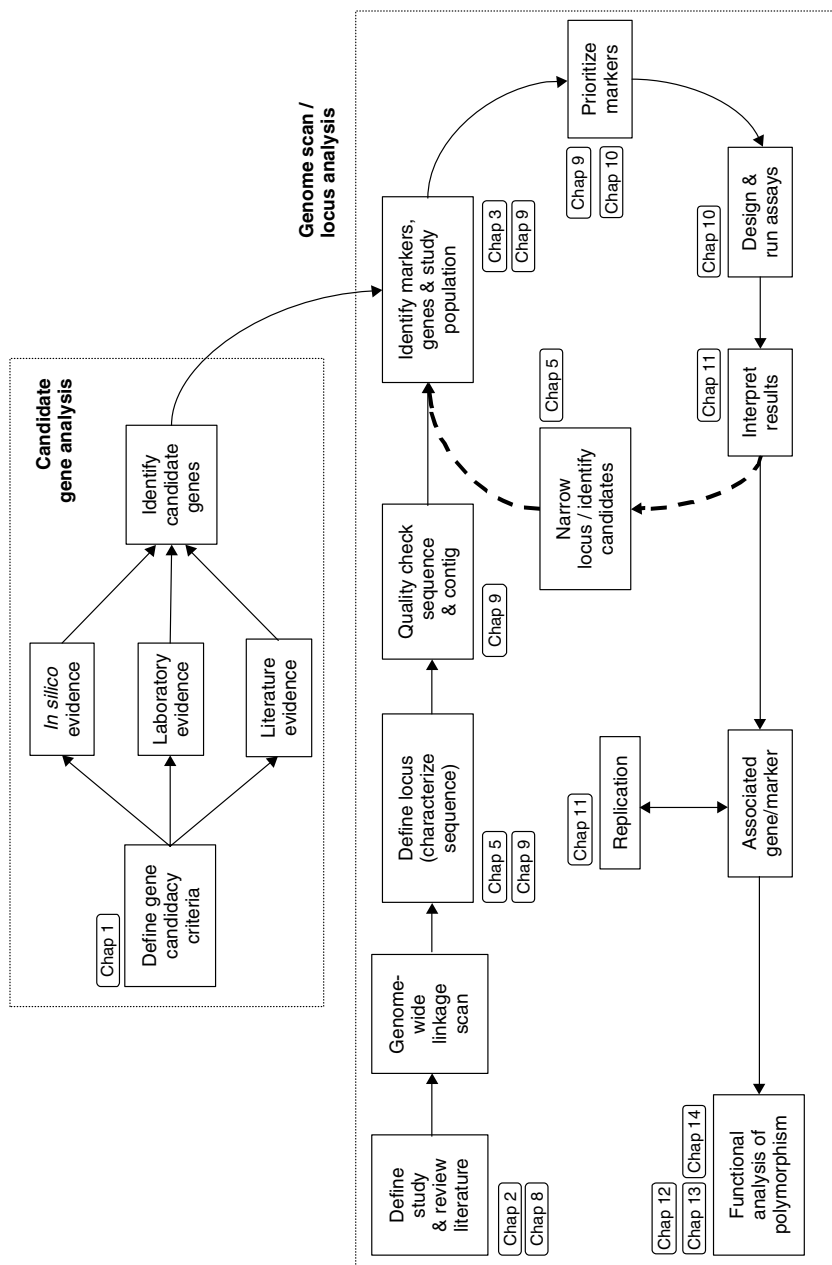


Figure 1.1 The genetic study process for complex disease, both candidate gene approaches and the follow-up of genome-wide linkage scans, highlighting chapters covering informatics aspects of each key step.

1.7 SELECTING CANDIDATE GENES FOR ANALYSIS

Candidate genes are typically selected for testing for association with a disease state on the basis of either (i) biological rationale; the gene encodes a product which the investigator has good reason to believe is involved in the disease process, (ii) the fact that the gene in question is located under a linkage peak, or (iii) both. The biggest problem with candidate gene analysis is that apparently excellent candidates are usually highly abundant and this surfeit of 'good' candidates is often difficult to rationalize.

Bioinformatics can be one of the most effective ways to help shorten, or more correctly prioritize, a candidate list without immediate and intensive laboratory follow-up. Firstly candidate criteria need to be determined based upon the phenotype in question. Detailed searches of the literature may help to flesh out knowledge of the disease and related pathways. Once a set of criteria is defined (for example which tissues are likely to be affected, which pathways are likely to be involved, and what types of genes are likely to mediate the observed phenotype), further literature review will help to 'round up the usual suspects', genes in known pathways with an established role in the phenotype under study. This is probably the most time-consuming step, but some tools can help to expedite this process, for example tools like OMIM can provide concise summaries of a disease area or gene family. Other databases encapsulate knowledge of pathways and regulatory networks, e.g. the *Kyoto Encyclopedia of Genes and Genomes* (KEGG; Kanehisa *et al.*, 2002). An alternative or parallel approach at this stage is to use a broader net to identify all genes which *could* be involved in the disease based on relaxed criteria such as tissue expression. Many *in silico* gene expression resources are available, including data derived from EST libraries, serial analysis of gene expression (SAGE; Velculescu *et al.*, 1995) data, microarray and RT-PCR data (see Chapter 15). For example, if the disease manifests in the lung, it is possible to distinguish genes that show lung expression from those that do not. This gives an opportunity to reduce emphasis on genes that show expression patterns which conflict with the disease hypothesis. However, it should be noted that electronic expression data is typically not comprehensive and care must be taken in using it to exclude the expression of a gene in a specific tissue. Low-level expression may not be detected by the method used; furthermore, gene expression may show temporal and spatial regulation — a gene may only be expressed during a specific phase of development or under particular conditions, e.g. cellular stress or differentiation.

1.8 PROGRESSING FROM CANDIDATE GENE TO DISEASE-SUSCEPTIBILITY GENE

In recent years, countless associations between genes and disease have been published, however many of these are likely to be spurious. Many reported associations show marginal *p*-values and subsequent studies often fail to replicate initial findings. Clearly *p*-values of around 0.05, generally accepted as the cut-off for a significant finding, will occur by chance for every 20 tests performed; this largely explains the general failure to reproduce promising primary results. However, real but very small effects giving marginal *p*-values are also difficult to replicate, leaving the investigator unsure as to the meaning of a failure to replicate. One approach for resolving the issue is to perform a rigorous meta-analysis using all available data, including both positive and negative associations. This type of analysis was recently used to demonstrate an association between the nuclear hormone receptor PPAR γ and diabetes, using data (previously regarded as equivocal)

drawn from a range of publications (Altshuler *et al.*, 2000b). Nonetheless, this approach relies on a lack of publication bias, i.e. the improbable assumption of an equal chance of publication for both positive and negative results.

Ultimately the biologist requires functional data to support an hypothetical genetic association; bioinformatics has a role to play here too. For example, DNA variants that alter subsequent amino-acid sequences can be checked for potential functional consequences using software tools (Chapters 12 and 14). Similarly, a thorough bioinformatic characterization of putative regulatory elements can give an indication of the possible impact of polymorphisms on *cis*-acting transcriptional motifs and the consequence on expression levels (Chapter 13). Bioinformatics can also assist in laboratory-based functional assessment of genes and polymorphisms; simple sequence manipulation tools coupled with genome sequence data can be used to design constructs for the *in vitro* and *in vivo* analysis of genes and polymorphisms using expression assays, transgenic mice and a host of other systems. However, perhaps the largest impact from bioinformatics on the field of functional characterization of genes will come from the development of powerful pattern recognition software for the identification of relationships between multitudes of transcripts analysed using microarrays. This approach has already proved useful in tumour classification by relating patterns of gene expression to response to chemotherapeutic agents (Butte *et al.*, 2000). An extension of this method should allow the elucidation of gene–gene interactions and the identification of common or converging biochemical pathways. Coupled with a knowledge of putative disease-related polymorphisms and comparable expression profiles in disease tissue, microarrays (together with the nascent field of proteomics; see Chapter 16) promise to be an extremely powerful future tool for the dissection of complex disease processes. Figure 1.2 illustrates approaches for gene characterization which are useful for both prioritizing candidate genes for analysis and establishing causality in a disease process. The chapter detailing each aspect is indicated.

1.9 COMPARATIVE GENETICS AND GENOMICS

We have already touched on the role of bioinformatics in relation to the identification of functionally important DNA motifs by cross-species comparison. This area is covered more fully in Chapters 9 and 12. Recently the sequencing of a number of genomes has been completed, including the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, the fruit fly *Drosophila melanogaster* and the nematode worm *Caenorhabditis elegans*; soon these will be joined by the puffer fish species *Fugu rubripes* and *Tetraodon nigroviridis*, the zebra fish *Danio rerio* and of course the mouse and rat. This has provided an unprecedented opportunity for large-scale genome comparisons, allowing researchers to make inferences not only with regard to the identification of conserved regulatory elements, but also about genome evolutionary dynamics. Whole genome availability also provides a complete platform for the design of *in vivo* paradigms of human disease, for example transgenic and gene knock-out animal models and more sophisticated spatially and temporally regulated conditional mutants.

Large-scale approaches to biochemical pathway dissection using expression microarrays in relatively simple organisms, particularly yeast, are also proving extremely promising. Whole genome expression profiles can be generated and correlated transcription profiles identified for related groups of genes. Coincident expression patterns are frequently indicative of subsequent protein–protein interactions and co-localization in protein complexes (Jansen *et al.*, 2002). Similar tissue-specific experiments can be performed for

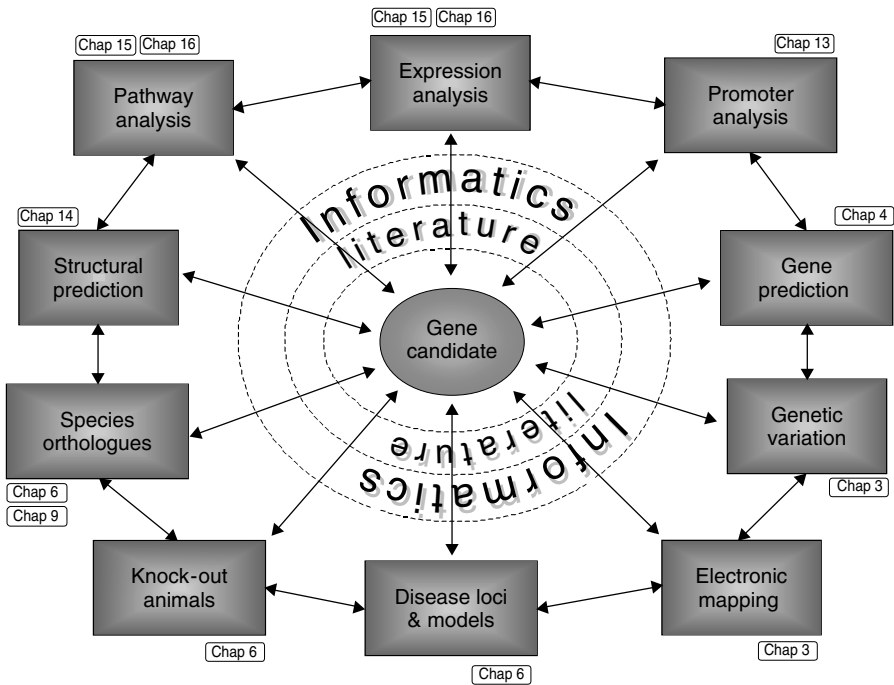


Figure 1.2 Approaches for gene characterization, indicating chapters detailing each aspect.

higher organisms, both for the purposes of identifying coincident transcription profiles for unravelling biochemical pathways and for comparison of diseased and normal tissues (see, for example Mody *et al.*, 2001; Saban *et al.*, 2001). Tissue derived from animal models such as mice can have advantages over using diseased human tissue: the disease model can be generated under a controlled environment, typically on an identical genetic background to the control tissue, and procurement of a significant number of high-quality tissue samples (essential for the extraction of good quality RNA) is more straightforward (see Chapter 15).

Thus far we have given a few examples of the impact of combining model organisms with high-throughput genomics technologies for improving our understanding of gene function and interaction, biochemical pathways and human disease (comparative genomics). Similar strides are being made in the field of comparative genetics (here we define genetics as phenotype-driven gene identification using genetic mapping procedures), particularly in the areas of mouse and rat genetics. The ability to perform controlled crosses such as inter-crosses and backcrosses (see Silver, 1995; Chapter 11) coupled with the development of fairly high density genetic maps over the last few years has rendered the mapping of monogenic traits in both mouse and rat a reasonably straightforward exercise. The impact of the completion of the mouse and rat genome sequences in the near future will be similar to the impact of the availability of the human genome on human genetics; indeed, the partially completed mouse and rat genomes are already giving significant improvements in speed of mapping and candidate gene

identification. These developments together with recently implemented large-scale mutagenesis programmes for the generation of monogenic mutants (see Chapter 6) promise to provide a significant increase in the mutant mouse resource in terms of simple disease models.

Significant progress has also been made in mapping complex traits in both the mouse and rat in recent years, including the development of software packages for the identification of quantitative trait loci (QTL; see Chapter 11). However, although experimental crosses can be designed to maximize the chances of success (unlike human studies), complex trait analysis in model organisms is still plagued by the difficulties in identifying and precisely localizing genes of relatively small effect. QTL linkage peaks are typically broad due to lack of absolute correspondence between genotype and phenotype and a consequent inability to identify unequivocal recombinant animals. In an attempt to overcome this limitation, mapping methods using ‘heterogenous stocks’ have recently been developed (Mott *et al.*, 2000). The heterogenous stock comprises a mouse line resulting from inter-crossing several different inbred strains and maintaining the resulting mixed stock through several generations (typically 30–60). Each chromosome from a mouse derived from a heterogenous stock consists of a mosaic of DNA from the different founding strains, allowing a fine mapping approach based on a knowledge of the ancestral alleles in the original inbred lines. Mott *et al.* have developed publicly available software for the analysis of heterogenous stocks (see Chapter 11).

Perhaps one of the most exciting developments in model organism genetics is the fusion of classical genetics with high-throughput genomics techniques. Microarrays provide a means of checking all genes within a QTL linkage peak for subtle differences in expression levels, potentially pinpointing the culprit gene. This tactic was used successfully to reveal the role of *Cd36* in metabolic defects, following linkage analysis in the rat (Aitman *et al.*, 1999). As an extension of this method, a gene expression profile may be treated as a quantitative trait and used as a phenotypic measure in linkage analysis for the identification of genes influencing the expression level, as a route to biochemical pathway expansion. Jansen and Nap (2001) recently coined the phrase ‘genetical genomics’ for this type of approach.

1.10 CONCLUSIONS

We hope this book will help the geneticist to design and complete more effective genetic analyses. Bioinformatics can have far-reaching effects on the way that a laboratory scientist works but obviously it will never entirely replace the laboratory process and is simply another set of tools to expedite the research of the practising biologist. Misconceptions regarding the power of bioinformatics as a stand-alone science are perhaps among the biggest mistakes that computer-based bioinformatics specialists can make and may even explain a degree of prejudice against bioinformatics—perceived by some as an ‘*in silico* science’ with little basis in reality. Taken to an extreme and without a balanced understanding of both the application of software tools and a good appreciation of basic biological principles, this is exactly what bioinformatics can be, but where bioinformatics proceeds as part of ‘wet’ and ‘dry’ cycles of investigation, both processes are stronger as a result. In this introduction we have briefly examined some of the experimental genetics processes which can be assisted by informatics; we now invite the reader to read on for more detailed coverage of each of these processes in the remaining chapters of this book.

REFERENCES

- Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PJ, Wahid FN, *et al.* (1999). Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nature Genet* **21**: 76–83.
- Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, *et al.* (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc Natl Acad Sci USA* **92**: 1684–1688.
- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, *et al.* (2000a). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.
- Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, *et al.* (2000b). The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet* **26**: 76–80.
- Antonellis A, Rogus JJ, Canani LH, Makita Y, Pezzolesi MG, Nam M, *et al.* (2002). A method for developing high-density SNP maps and its application at the Type 1 Angiotensin II Receptor (AGTR1) Locus. *Genomics* **79**: 326–332.
- Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci USA* **97**: 12182–12186.
- Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, *et al.* (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**: 1423–1427.
- Collins A, Lonjou C, Morton NE. (1999). Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci USA* **96**: 15173–15177.
- Debouck C, Metcalf B. (2000). The impact of genomics on drug discovery. *Ann Rev Pharmacol Toxicol* **40**: 193–207.
- Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. (2002). Hobbies: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res* **30**: 387–391.
- Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, *et al.* (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**: 599–603.
- Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971–983.
- Jansen RC, Nap JP. (2001). Genetical genomics: the added value from segregation. *Trends Genet* **17**: 388–391.
- Jansen R, Greenbaum D, Gerstein M. (2002). Relating whole-genome expression data with protein–protein interactions. *Genome Res* **12**: 37–46.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.
- Kanehisa M, Goto S, Kawashima S, Nakaya A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**: 42–46.
- Kruglyak L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet* **22**: 139–144.
- Kruglyak L, Nickerson DA. (2001). Variation is the spice of life. *Nature Genet* **27**: 234–236.

- Lai E, Riley J, Purvis I, Roses A. (1998). A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**: 31–38.
- Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, *et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**: 66–71.
- Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, *et al.* (2001). Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc Natl Acad Sci USA* **98**: 8862–8867.
- Mott R, Talbot CJ, Turri MG, Collins AC, Flint J. (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci USA* **97**: 12649–12654.
- Pericak-Vance MA, Bass ML, Yamaoka LH, Gaskell PC, Scott WK, Terwedow HA, *et al.* (1998). Complete genomic screen in late-onset familial Alzheimer's disease. *Neurobiol Aging* **19** (1 Suppl): S39–S42.
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, *et al.* (2001). Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.
- Risch NJ. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**: 847–856.
- Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, *et al.* (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**: 1066–1073.
- Saban MR, Hellmich H, Nguyen NB, Winston J, Hammond TG, Saban R. (2001). Time course of LPS-induced gene expression in a mouse model of genitourinary inflammation. *Physiol Genomics* **5**: 147–160.
- Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, *et al.* (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci USA* **90**: 1977–1981.
- Saunders AM, Strittmatter WJ, Schmechel D, St. George-Hyslop PH, Pericak-Vance MA, Joo SH, *et al.* (1993a). Association of apolipoprotein E allele E4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**: 1467–1472.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.
- Silver LM. (1995). *Mouse Genetics: Concepts and Applications*. Oxford University Press: Oxford, UK.
- Spielman RS, McGinnis RE, Ewen WJ (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506–516.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. (1995). Serial analysis of gene expression. *Science* **270**: 484–487.