PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

An overview of recent developments in genomics and associated statistical methods

Peter J. Bickel, James B. Brown, Haiyan Huang and Qunhua Li

Phil. Trans. R. Soc. A 2009 **367**, 4313-4337
doi: 10.1098/rsta.2009.0164

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 4313–4337
doi:10.1098/rsta.2009.0164

Review

An overview of recent developments in genomics and associated statistical methods

By Peter J. Bickel[1], James B. Brown[2], Haiyan Huang[1,*] and Qunhua Li[1]

[1]*Department of Statistics, and* [2]*Applied Science & Technology, University of California, Berkeley, CA, USA*

The landscape of genomics has changed drastically in the last two decades. Increasingly inexpensive sequencing has shifted the primary focus from the acquisition of biological sequences to the study of biological function. Assays have been developed to study many intricacies of biological systems, and publicly available databases have given rise to integrative analyses that combine information from many sources to draw complex conclusions. Such research was the focus of the recent workshop at the Isaac Newton Institute, 'High dimensional statistics in biology'. Many computational methods from modern genomics and related disciplines were presented and discussed. Using, as much as possible, the material from these talks, we give an overview of modern genomics: from the essential assays that make data-generation possible, to the statistical methods that yield meaningful inference. We point to current analytical challenges, where novel methods, or novel applications of extant methods, are presently needed.

Keywords: genomics; statistical genomics; experimental biotechnologies; high-dimensional bio-data analysis

1. Introduction

The central dogma of molecular biology, as enunciated by Crick (1958), specified the instruction manual, DNA (encoding genes), and that genes were transcribed into RNA to ultimately produce the basic operational elements of cellular biology, proteins, whose interactions, through many levels of complexity, result in functioning living cells. This was the first description of the action of genes. After an enormous experimental effort spanning the last half century, made possible by the development of many assays and technological advances in computing, sensing and imaging, it has become apparent that the basic instruction manual and its processing are vastly more sophisticated than what was imagined in the 1950s. Genes were found to account for at most 2 per cent of the human

*Author for correspondence (hhuang@stat.berkeley.edu).
All authors contributed equally to this work.

One contribution of 11 to a Theme Issue 'Statistical challenges of high-dimensional data'.

genome's string of 3 billion base pairs. The remaining 'non-coding' portion, initially labelled as 'junk DNA', is responsible for regulation of the coding sequence and self-regulation via a list of mechanisms that continues to grow each year.

Remarkable technologies such as microarrays and their descendants, high-throughput sequencing, *in vivo* imaging techniques and many others have enabled biologists to begin to analyse function at molecular and higher scales. The various aspects of these analyses have coalesced as 'omics': transcriptomics, the study of gene–gene regulation, in particular, DNA–protein interactions; proteomics, the large-scale study of the structures and functions of proteins; metabolomics, the study of small-molecule metabolite profiles. These processes are tightly linked and the utility of these labels is unclear; see Greenbaum *et al.* (2001) for an amusing discussion. Genomics, which encompasses all these, can be thought of as the science linking DNA structures with functions at the cellular level.

The primary focus of modern biology is to link genotype to phenotype, interpreted broadly, from the level of the cellular environment to links with development and disease. A common feature of these activities is the generation of enormous amounts of complex data, which, as is common in science, though gathered for the study of one group of questions, can be fruitfully integrated with other types of data to answer additional questions. Since all biological data tend to be noisy, statistical models and methods are a key element of analysis.

The purpose of this paper is to give an overview of current statistical applications in genomics, primarily the study of genomes at the DNA and mRNA (transcription) levels. The occasion initially prompting this article was the recent workshop on high dimensional statistics and biology held at the Newton Institute in Cambridge, 31 March–4 April 2008. We will largely use the papers and content presented at the workshop as illustrative examples. We are not so foolhardy as to attempt to deal with the broad interaction of biology and statistics.

This paper is organized as follows. In §2, we outline the historical development of genomics and its bases in genetics and biochemistry. In §3, we introduce various types of modern biological technologies, the data being generated and the biological questions that are being posed. In §4, we summarize the methods of analysis that have been developed and indicate possible weaknesses as well as methods in the literature that may meet these challenges better. In §5, we discuss possible new directions of research and point to where new analyses and tools may be needed.

2. A brief history of genomics

Darwin published *On the Origin of Species* in 1859, outlining the process of natural selection (Darwin 1900). Contemporary with Darwin's work, G. Mendel was in the process of ascertaining the first statistical theory of inheritance (Weiling 1991). Genetics can be said to have begun when Mendel coined the term, 'factors', to describe the, then unseen, means of conveyance by which traits were transmitted from generation to generation. During the early twentieth century, mathematical scientists, in particular, R. A. Fisher, J. B. S. Haldane and S. Wright, assembled the algebraic analysis of Mendelian inheritance, developed the statistical framework of population genetics, and so infused the theory of evolution with genetic explanations and corresponding statistical models (Fisher 1930; Wright 1930; Haldane 1932; Griffiths *et al.* 2000).

In the 1940s and early 1950s, the biological focus of investigations shifted to the physical nature of the gene. In 1944, DNA was successfully isolated by Oswald Avery as the basic genetic material (Avery *et al.* 1944). Watson & Crick (1953) discovered the double helical structure of double-stranded DNA, and the relation between its structure and capacity to store and transmit information. These, and many other discoveries, marked the transition from genetics at the level of organisms to molecular genetics. In 1958, Crick (1970) first enunciated the central dogma of molecular biology: DNA codes for RNA, which codes for protein. The regulation of gene expression then became a central issue throughout the 1960s.

Since the 1970s, technologies for sequencing DNA, RNA and proteins made possible the direct study of genetic material, and molecular biology entered the genomic era. In 1972, W. Fiers determined the sequence of a bacterial gene (Min Jou *et al.* 1972). In 1977, Sanger first sequenced the complete genomes of a virus and a mitochondrion (Sanger *et al.* 1977), and later established systematic protocols for sequencing, genome mapping and sequence analyses. Many studies were significantly enhanced by these developments. In the last decades of the twentieth century, bioinformatics research matured rapidly as a field, driven by advances in sequencing technology, as well as computer hardware with which to analyse mounting stores of data.

During the 1980s and 1990s, the polymerase chain reaction (PCR; Bartlett & Stirling 2003), automated DNA sequencing, and microarrays (Kulesh *et al.* 1987; Schena *et al.* 1995) solidified genomics as a pre-eminent discipline within the life sciences. These modern assays enabled biologists to resolve questions at a scale and depth that were not previously possible, e.g. determining entire eukaryotic genomes, constructing fine-scale genome-wide genetic maps and analysing and integrating various genomic, proteomic and functional information to elucidate gene-regulatory networks.

The first collaboration of massive scope was the Human Genome Project, and it involved contributions from over 100 laboratories (Francis *et al.* 2003). It was initiated in 1990 with the goal of 'mapping' the entire human genome. In April 2003, 13 years after its inception, the successful completion of the Human Genome Project was reported, with 99 per cent of the genome sequenced to 99.99 per cent accuracy. A parallel competitive project was conducted by Celera Genomics, a private company (Venter *et al.* 2001). Many more genomes have been sequenced. As of February 2009, sequences from around 250 000 organisms were publicly available (http://www.ncbi.nlm.nih.gov).

With the completion of sequencing projects for many model organisms, molecular biology entered a new era. The focus of research has turned from the determination of sequences and the genetic units of inheritance to the systematic study of complex interactions of structure and function in biological systems, sometimes called systems biology. Since functions do not stop at the cellular level, this can go beyond genomics as we have defined it.

The ENCODE project (The ENCODE Project Consortium 2004), begun in September 2003, is a large scale example of systems biology. A necessary prerequisite of this project is to define 'function' in genomics. Prior to the 1970s, a gene was defined by promoter sequence upstream of a 'cystron', a contiguous, transcribed unit coding for protein. This simplistic view was refined by the discovery of exons and introns in 1977 (Gilbert 1978). The ENCODE Project Consortium (2007) reported, among other things, that many genes, far more than previously established, generated chimeric transcripts: single RNA transcripts which come from two or more neighbouring genes. If these chimeric elements turn out to have important biological functions, then our notion of a gene may be redefined yet again.

Next-generation DNA sequencers are the most recently developed, game-changing technology (Mardis 2008). They are capable of sequencing billions of base pairs of DNA in each automated run, producing more data than any other technology in the history of biology. These new platforms have inspired diverse uses. Besides applications in classical genomic studies, cheap and rapid sequencing may revolutionize diagnostic medicine by permitting an unprecedented degree of hyper-differentiation in health-care practices. Indeed, individual genetic profiling has been brought to the medical domain. It is expected that this will eventually permit more precise dosage control and drug choice. The 1000 Genome Project will soon release the sequence of 1000 individual human genomes.

Clearly, the boundaries between genomics and other parts of biology are disappearing. In fact, several of the talks at the workshop followed these larger aims. As in other fields, the rapidly evolving and diversifying fields of biological research, coupled with technological advances, have given rise to needs for novel computational or analytical techniques, e.g. tools for systematic or integrative analysis of high-dimensional, diverse biological data. Studies dependent upon the integration of multiple data-types are becoming increasingly prevalent, especially in large-scale collaborations such as the ENCODE project. A wide variety of quantitative scientists (computational biologists, statisticians, computer scientists, physicists, biochemists, etc.) are working to create, refine and test computational models to integrate various data-types, but many more are needed and welcome.

3. Basic technologies

In this section, we briefly describe some fundamental experimental methods in genomic research and the types of variability associated with each method. Our goal is to lay out the sources of analytical challenges in genomic research, thus motivating the statistical methods in the next section. As we shall see in this section, basic methods have been combined in groups, reminiscent of the combinatorial complexity of genomes, to generate further, higher levels of data. The results of thousands of assays have been collected in databases such as NCBI (www.ncbi.nlm.nih.gov) and the Ensembl genome browser (www.ensembl.org) enabling their results to be portrayed as 'features', annotations defined across the genome in genomic coordinates. These databases are enabling us to begin to trace the steps from genome to regulome, to proteome, to metabolome and, with great gaps in our knowledge, the fundamental interactions of genotype and environment that lead to observable phenotypes and disease.

(*a*) *Gel electrophoresis*

Introduced in the 1950s, gel electrophoresis is a method for separating molecules according to their rate of traversal across an electrified gel (Macinnes 1960). This technique can be used to determine the lengths of molecular fragments down to the single basepair scale, a property employed in a variety of DNA sequencing protocols.

(*b*) *Blotting*

The Southern Blot assay, developed in the 1970s, is an inexpensive and rapid means of determining the presence or absence of a particular DNA sequence in a large pool of unknown fragments (Southern 1975). The Northern Blot (Bor *et al.* 2006) and the immunoblot (or Western Blot; Towbin *et al.* 1979) followed within a few years. In the 1990s a 'Far-Eastern Blot' was developed for the analysis of lipids (Ishikawa & Taki 1998). Although initially used qualitatively, blotting can also be used quantitatively in conjunction with the other assays. Early versions of microarray, described later, evolved from this technology.

(*c*) *Polymerase chain reaction* (*PCR*)

PCR is a chemical process that exponentially amplifies the number of copies of segments of DNA, up to tens of kilobases in length (Tindall & Kunkel 1988). PCR plays a key role in producing most types of data, especially in sequencing. Amplification errors in PCR are rare, and in modern protocols are at worst of the order of 1 base pair in 9000. So in most applications the statistical issues are minor (Tindall & Kunkel 1988). However, when thousands, millions or even billions of different DNA sequences are being simultaneously amplified in the same reaction, sometimes called 'multiplex ligation-dependent probe amplification' (Schouten *et al.* 2002), it may be that there exist subtle differences in the rates of amplification of the sequences and hence may call for more complex statistical modelling.

(*d*) *Microarrays*

A microarray, introduced in 1995, is a chip 'printed' with thousands of variants of a particular polymer, such as DNA, RNA or cDNA (Schena *et al.* 1995). Each variant appears in a tiny dot, known as a probe or spot, containing many copies of the same sequence. The target or sample sequences, labelled with fluorescence tags, are washed across the chip. The dots, when bound by the target sequences, either fluoresce or are detected by some other imaging method. The luminescent intensity of each probe indicates the relative quantity of the target present. A recent important application of microarrays is the 'tiling array', where nearly every sequence in a genome, of at least some particular length, e.g. 30 bp, is tiled on the array. Tiling arrays are powerful for genome-wide investigations, and have been used in many studies, such as transcriptome mapping and protein-binding sites mapping (Bulyk 2006; Sabo *et al.* 2006). Though different in wet-laboratory protocols, various applications of microarrays end up with the same output format, i.e. a high-resolution image of thousands of variously illuminated little dots. The worth of the assay is predicated on image processing and subsequent statistical analysis.

Microarrays exemplify the typical paradigms of modern statistics: observations with thousands of dimensions are repeated only a few times under possibly different conditions; there exists high sparsity and noise in data. Noise enters from many sources: non-specific binding of target to probe (false positives), unanticipated inaccessibility of certain probes (false negatives), and unknown relationship (generally nonlinear) between total measured luminescence and quantity of the target (Yang *et al.* 2002; Dudoit *et al.* 2003; Rozowsky *et al.* 2005). A rich statistical literature tackling these challenges has emerged, such as multiple testing (Dudoit *et al.* 2003), false discovery rates and related measures (Benjamini & Hochberg 1995; Storey 2002; Storey & Tibshirani 2003; Benjamini 2008).

(*e*) *DNA sequencing*

Sequencing is a key technique in genomics. Since the 1970s, a wide variety of DNA sequencing technologies have been developed resulting in gains of many orders of magnitude in speed and accuracy (Sanger *et al.* 1977; Smith *et al.* 1986; Hall 2007). The Sanger method (Sanger *et al.* 1977), or chain-terminator method, has been dominant in the field, especially at the time of the Human Genome Project. This method relies on replicating many copies of the 'to-be-sequenced' DNA segment in four separate *in vitro* reactions, where each reaction is associated with one of the four dideoxynucleotides (A's, C's, G's and T's). Whenever a dideoxynucleotide instead of a deoxynucleotide is recruited, the replication then suddenly terminates at precisely that base. Hence, each reaction produces many DNA fragments of various lengths with the same, known base at the end of each fragment. Next, using gel electrophoresis on these fragments, with one separate lane of gel for each of the four reactions, the to-be-sequenced DNA sequence can simply be read from the length distribution on the gel. Later developments modified this protocol to use fluorescently labelled dideoxynucleotides, which permits one reaction containing all four chain terminating bases, rather than four separate reactions (Smith *et al.* 1986). Recently, next-generation sequencing technologies, based on parallelizing the sequencing process, have made it possible to produce thousands or millions of sequences at a cost much lower than those of the traditional terminator-based methods (Hall 2007).

Arguably, the most important application of sequencing in the last two decades has been to sequence the model genomes (Francis *et al.* 2003). It consists of two basic steps. First, many copies of subsequences of a few hundreds of basepairs are generated by the sequencers; these small pieces of sequences with every base known are called as 'reads'. Then the reads are assembled to reconstruct the genome. Primary computational difficulties come from both steps. At the read-generation step efficient methods/algorithms are needed to analyse image data and correctly call a base at each position. At the assembly step there is no prior knowledge of how the reads are supposed to fit together, and furthermore, there are possible base-calling errors. Assembly is analogous to solving a massive one-dimensional jigsaw puzzle with the pathological property that many of the pieces occur many times in different places due to the repetitive nature of genomes. Fairly sophisticated mathematical models and a variety of 'base-calling', 'mapping' or 'assembly' algorithms have been developed to solve the problems with varied degrees of success (Ewing & Green 1998; Venter *et al.* 2001; Bentley 2006; Zerbino & Birney 2008).

Sequencing technologies also play a key role in the assays that require the identification of DNA sequences. Examples of such use include the serial analysis of gene expression (Velculescu *et al.* 1995) and its updated version, the cap analysis of gene expression (Shiraki *et al.* 2003). Both methods measure the relative frequency of different RNA transcripts in a high-throughput fashion. The RNA-seq assay, relying on the next-generation sequencing instruments, provides more comprehensive information on a sample's RNA contents (known as transcriptome data; Toth & Biggin 2000). Another noteworthy example is the ChIP-seq assay, which combines next-generation sequencing with ChIP (chromatin immunoprecipitation; described in the following subsection) for detecting the genome-wide DNA binding sites of a protein of interest (Johnson *et al.* 2007). Noise of these data comes from possible sequencing (or base calling) errors and the need to map the sequenced segments back to the genome (mapping is often not unique). Several methods have been developed to analyse these data using ad hoc approaches or some probabilistic modelling (Johnson *et al.* 2007; Boyle *et al.* 2008; Jothi *et al.* 2008; Valouev *et al.* 2008; Zhang *et al.* 2008; Rozowsky *et al.* 2009).

(*f*) *Chromatin immunoprecipitation assay*

Chromatin immunoprecipitation (ChIP) is a popular approach for measuring the binding of proteins to DNA in living cells (Gilmour & Lis 1987). It involves fragmenting chromatin (the protein-DNA complex that is the natural state of chromosomes) into small fragments, and then precipitating fragments that include a particular protein of interest. This is done using an antibody specific to that protein. ChIP has been generally combined with microarrays (known as ChIP-chip) or sequencing technology (known as ChIP-seq; Toth & Biggin 2000) to perform genome-wide identification of binding regions such as transcription factor binding sites or methylation sites (Dekker *et al.* 2002).

Noise can enter the assay due to the protein of interest being cross-linked to a genomic region that it does not directly contact. This can occur via intermediary proteins, where protein chains form bridges to other genomic regions, or due to the cross reaction of the antibody with another protein (although this latter source of noise is usually detected during the assay preparation). Statistical techniques have been and are being developed to separate the signal (e.g. specific binding events, chimeric sequences, methylation patterns) from the noise (e.g. non-specific binding, misleading cross-linking, sequencing or microarray errors; Schena *et al.* 1995; Rozowsky *et al.* 2005, 2009; Sabo *et al.* 2006). Since such studies are now being carried out on the scale of whole genomes, methods such as the false discovery rate (FDR) also play an important role (Benjamini 2008).

(*g*) *Sequence alignment*

The central aim in early sequence analysis was the identification of homologous sequences. Biosequences are said to be homologous, and therefore likely to share common function, if they are descendants of a common ancestral sequence

(Kimura 1983). Since sequence homology is generally concluded on the basis of sequence similarity, this aim gives rise to the need for computational methods and algorithms, particularly for the discovery of duplicated sequences (paralogs) in the same genome and highly similar sequences (orthologs) in the genomes of related species (Burge & Karlin 1998).

Early efforts in sequence comparison were to align amino acid sequences, which are generally short (at most several thousand residues). The Needleman–Wunsch algorithm, developed in the 1970s based on dynamic programming, can give a 'global alignment' of two sequences optimal under a particular, user-defined, substitution and insertion/deletion matrix (Needleman & Wunsch 1970). The Smith–Waterman algorithm generalizes Needleman–Wunsch by finding optimal 'local alignments' of subsequences within longer molecules (Smith & Waterman 1981). Both of these time-intensive algorithms were precursors to high-throughput technologies capable of comparing millions of sequences. The first truly high-throughput tool was the basic local alignment search tool (BLAST) for searching a query sequence against a database to detect all elements in the database homologous to at least a subsequence of the query sequence (Altschul *et al.* 1990). Statistical evaluation of the searching results has been based on the Karlin–Altschul statistics (Karlin & Altschul 1990). BLAT, the BLAST-like alignment tool, was introduced in the 1990s as essentially a faster and more sensitive version of the original (Kent 2002).

Tools such as BLAST have made pair-wise alignment quite fast. However, multiple alignment, an increasingly popular aim, is computationally far more expensive. Most multiple alignment approaches fall into a class known as 'progressive alignment' (Brudno *et al.* 2003). Its underlying strategy is as follows: many pair-wise alignments are conducted, and those pair-wise alignments are aggregated. Aggregation is non-trivial due to the size of the set of sequences under consideration, and because the particular technique varies among methods of multiple alignment (Blanchette *et al.* 2004; Blanchette 2007). The problem here tends to be computational rather than statistical, but remains very challenging.

We have chosen to highlight several of the many canonical technologies presently in use in molecular biology. These and other technologies and techniques come together in pairs and higher-order groupings to generate the diversity of biological data presently being produced in laboratories around the world. As we pointed out, ChIP, a selection protocol, can be combined with any of a number of sensing technologies, including sequencing, microarrays or PCR, depending upon the desired scale and resolution of the assay. Such combinations of assays extend beyond combining biochemical selection and sensing procedures. Entirely diverse experiments can be and are regularly combined in mathematical or statistical models to permit inferences that cannot be tested directly. The integration of many techniques from biochemistry and statistics is likely to be the direction of molecular and systems biology in the next decade (Segal *et al.* 2008).

The intrinsically high-dimensional nature of biological data will continue to provide novel challenges for both statisticians and computer scientists in the coming decades. Presently, it is often necessary to make sacrifices in statistical methodology in order to develop computationally tractable models. In the following section, we discuss the extant statistical and mathematical approaches that have thus far facilitated studies, and point out areas where new methods or applications are needed.

4. Data analysis

As discussed throughout this paper, the advance of technology has drastically broadened the scope of biological research. Research interests have diversified with the technical capacity to investigate them, and now vary from genealogy to mapping the three-dimensional structure of chromatin in living cells. Contemporary questions are as specific as 'Does ascorbic acid inhibit nitrosamines?' or as exploratory as 'Can we classify the functional features of non-coding elements evolutionarily conserved across mammalian species?' The approaches used to address such issues can be qualitative or quantitative, and often vary across levels of complexity.

In classical hypothesis-driven research, the biologist seeks to test putative actions or interactions via experiments with explicit incentives, e.g. knock-out assays are used to investigate the functional role of genes. Such experiments produce the basic data and information needed for validation. As in any field, the extent to which a hypothesis can be directly addressed is determined by the effectiveness of experimental design, the power of the applied technology and the capacity of analysts to interpret the output. For instance, to study the gene function, one may imagine that in an ideal world, one would simply record a 'movie' of the relevant segment of chromatin, and watch the unfolding process of transcription, translation and the downstream action of the folded protein. This is, of course, presently impossible, and instead for such investigations biologists rely upon a variety of technologies, including the above-mentioned knock-out assays, to produce a host of data types, each with their own idiosyncrasies and signal-to-noise ratios. Technologies and assays have been born of a process of iterative refinement, where wet-laboratory biology is progressively informed by the challenges of data analysis. The ChIP-seq experiment is an example of such an ongoing interaction. What constitutes an appropriate negative control, as well as the process by which any negative control should be applied to the assay signal, has yet to be determined, and will require input both from organic chemists and statistical analysts.

Tukey in his famous 1962 paper (Tukey 1962) describes data analysis by part of what it contains: 'Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole. Large parts of data analysis are incisive, laying bare indications which we could not perceive by simple and direct examination of the raw data, but these too are only parts, not the whole. Some parts of data analysis, as the term is here stretched beyond its philology, are allocation, in the sense that they guide us in the distribution of effort and other valuable considerations in observation, experimentation or analysis. Data analysis is a larger and more varied field than inference or incisive procedures or allocation.' We briefly touch on subsets of these themes under their more modern rubrics, 'exploratory' (also a Tukeyism) for incisive, 'validatory' for inference, 'experimental design' under allocation, and 'prediction' for the parts now referred to as machine learning.

Exploratory analysis is essentially visual. This can mean trivial transformations, such as looking at data using histograms or boxplots, or examining data-derived quantities like regression residual plots or correlation coefficients. Or, it can mean more sophisticated techniques: dimension reduction using principal component analysis (PCA; Alter *et al.* 2003); low-dimensional projections of high-dimensional data; or cluster analysis. The goal is to reveal features invisible in the original data. Validatory analysis corresponds to using tools such as hypothesis testing and confidence regions to determine if features found by exploratory analysis are simply due to chance. Experimental design, which precedes data collection, ensures that the data gathered are as informative as possible under cost constraints. For instance, for gene-expression microarray assays, the number of biological and technical replicates has to be chosen so that variability between gene-expression levels is not washed out by intrinsic variability due to biological and technical sources.

An aspect not explicitly mentioned in Tukey's description that we will dwell on extensively in the following is probabilistic modelling. It is sometimes the case that probabilistic models of how the data are generated precede exploratory analysis and are partly based on physical considerations. An example would be the formula for the binding affinity constant in a reaction involving reagents A and B in thermodynamics,

$$K_A = \frac{[\text{fraction bound}]}{[\text{free } A][\text{free } B]}. \tag{4.1}$$

This is a basic element of probabilistic models for binding of a transcription factor (protein) to a given oligonucleotide (short sequence of DNA), a primary mechanism of gene transcriptional regulation (Djordjevic *et al.* 2003).

It is only after we have a probabilistic model of the data that we can talk about validatory analysis. Technically, constructing a model based on exploratory analysis and fitting it on the *same* data makes validatory statements somewhat questionable and, in the context of prediction, is dangerous, as we argue below. But, in practice, it is always done since we usually do not know enough about biological phenomena to postulate hard and fast models. The expectation is that ultimate validation will come from new data or other evidence. The danger of postulating a model based on poor prior information is much greater, since all validatory statements depend on the validity of the model.

The last aspect we believe needs to be added is prediction, the main concern of machine learning. In one of the main types of prediction problems, classification, we wish to predict a yet-to-be determined outcome, e.g. whether an individual has cancer or not based on features of his or her genotype. We do this using a prediction rule based on a training sample of individuals whose genotype and disease status are known. Clearly, real validation here is only obtainable by ascertaining the individual's true disease status. We can, however, try to estimate the probability of misclassification in advance. If we do this naively by simply counting incorrect decisions using the rule on the training sample used to fit it we will underestimate possibly grossly and generate consequences such as selection bias (Ambroise & McLachlan 2002). For studies involving modern high-throughput technologies (e.g. microarrays), an issue that is always present and has become paramount is speed of computation. We will discuss this as we go along.

We now turn to discussion of subtopics under these broad headings, with illustrative examples from the workshop talks and other papers.

(*a*) *Exploratory data analysis*

(i) *Clustering*

Clustering is of particular importance given that dataset dimension is well beyond visualization. Many applications of clustering were discussed in the workshop, including classification of differentiation of stem cells (Bertone), cell types from microscopy data (Huber) and different virus types (Beerenwinkel).

The goal can crudely be defined as grouping the like and separating the unlike. However what 'like' means depends on the definition of likeness. If the points to be clustered are in a Euclidean space then it is natural to use distance between points as a measure of similarity, at least if the features (coordinates) are on the same scale. This is the type of problem treated by most types of clustering. Included in this approach are many clustering techniques, such as the classical agglomerative and other hierarchical methods, k-means clustering and other disaggregative methods. The classic statistical approach is to use mixture models, used and discussed by McLachlan *et al.* (2008), for clustering genes with different expression patterns using microarray experiments. An excellent reference for all the above methods is Hartigan's 1975 book (Hartigan 1975).

A key element in analysing high-dimensional and noisy biological signals is dimension reduction preceding clustering, for instance PCA (Alter *et al.* 2003), which will be discussed further later. Briefly, the empirical covariance matrix of the n points to be clustered is formed and the basis given by the 2 or 3 eigenvectors corresponding to the largest eigenvalues is used to give a representation in which cluster membership may be easy to identify visually. In biology, metrics other than Euclidean distance are often needed. A canonical example is the creation of gene clusters based on their expression in series of microarray experiments, where metrics such as those based on putting expression scores on the same scale are used. Often, numerical similarity measures between vectors of features which are not necessarily numerical are given in $n \times n$ matrix form. An example is the cosine measure to compare phenotype similarity (Lage *et al.* 2007). If the resulting matrix is positive semi-definite, the vectors can be identified with functions in a reproducible kernel Hilbert space and it may be appropriate to base a method of clustering on the eigenvectors and eigenvalues of the (normalized) similarity matrix as a generalization of PCA (Lanckriet *et al.* 2004). A huge literature on clustering using spectral properties of appropriate matrices has developed, in particular with so-called graph clustering. The relationships of these methods to natural properties of random walks and diffusions on appropriate spaces have been well explored (Shi & Malik 2000; Meila & Shi 2001; Ng *et al.* 2002; Nadler *et al.* 2005). These methods have only started appearing in the biological literature but are becoming more appreciated given that they provide natural methods of dimension reduction as well as clustering.

(*b*) *Prediction*

In the prediction literature, classification is called supervised learning while clustering is known as unsupervised, the difference being the presence of a training sample. There are many types of classification methods. Among the most popular are neural nets, logistic regression, classification and regression trees (CART), support vector machines and k-nearest-neighbour rules. The ones judged most effective currently are boosting and random forests. All such methods are reviewed from a practical point of view in the book by Hastie *et al.* (2009).

The basic principle behind all these methods is the same. Given a training sample of feature vectors and known outcomes, $(X_1, Y_1), \ldots, (X_n, Y_n)$, we wish to construct a rule which given a new X predicts its yet-to-be determined Y as well as possible. Here, for Y, several examples can be found in the workshop lectures: Y can be a disease state in the talk by Huang, a protein complex in the talk by Brunak, an mi-RNA gene in the talk by Enright. In classification the number of possible values of outcomes Y is finite. In the Brunak talk, the number of outcomes was given by the number of possible protein complexes. The feature vector X can consist of quantifications of gene expression, as in the Huang talk. If the training sample were infinite and not subject to selection bias, there would be a unique form of best method, the Bayes rule, deciding which value of Y is most likely given the observed X. In practice, one often has a small number of samples in relation to the dimension of X. The methods mentioned implicitly estimate these likelihood ratios, though often this is far from apparent. For instance, CART and random forests build decision trees, while the k-nearest-neighbour rule classifies the new X according to the majority category (value) of Y among the training set X's which are the k nearest neighbours of X in an appropriate metric. The major issue if X is high-dimensional is the problem of overfitting; the rule does a superb job on the training sample, but is poor on a new X. The simplest example of this phenomenon is the k-nearest-neighbour rule, which predicts perfectly in the training sample, but no matter how large the training sample is, does not approximate the Bayes rule. These issues appeared in Buhlmann's talk at the workshop and will be discussed further below. In this context, he modified a method for regression analysis between expression levels and motif occurrence frequencies (Conlon *et al.* 2003), avoiding overfitting by using variable selection.

(*c*) *Probabilistic models*

As we have noted, it is now common to face problems in which hundreds or thousands of elements are linked in complex ways and complementary information is shared between different data sources or different types. For example, in complex diseases, phenotypes are often determined not by a single or just few genes, but by the underlying structure of genetic pathways that involve many genes. Probabilistic models are an excellent way of organizing our thinking in such situations. We necessarily want to make them reflect as much subject-matter knowledge in terms of the data-gathering methods and known biology as possible.

Probabilistic models feature in exploratory analysis, predictive and validatory aspects of statistics. Their use in exploratory analysis is essentially implicit and it is often unclear whether the exploratory tool preceded or followed from the model. Gauss introduced the method of least squares, as opposed to the method of least absolute deviations favoured by Laplace, for computational reasons and proposed the Gaussian distribution because it was the one for which least squares estimates agreed with maximum likelihood (Stigler 1986). In prediction, probability models also play an implicit role since validation of predictions should be external to the data used to predict and hence model-free. They are, as we shall see, used primarily for predictive purposes in genomics.

(i) *Regression models*

In biology, the dimension of covariates derived from experiments is a major issue. This is well illustrated in Buhlmann's talk. His lecture introduces a special case of a class of models which have been used throughout the sciences and which we now describe. The situation that the number of potential explanatory variables substantially exceeds the number of observations (known as the large p small n problem) is prevalent in high-throughput data analysis, not just in genomics. A useful and frequently used model for problems of this kind is the regression model, with the simplest form written as

$$Y = X\beta + \varepsilon, \tag{4.2}$$

where X is an $m \times p$ matrix of the values of the p predictive variables associated with each of the n observations and ε is a 'noise' vector. It is expected that the column vector β will be sparse. For instance, if Y comes from measures relating to phenotype and X is the matrix of gene expression scores coming from a microarray, most genes will have no bearing on Y. Though the form of the model itself is simple, the potential high noise and combinatorial complexity of the problem (2^p subsets with p features) impose challenges. To make effective predictions and select important variables, various regularization methods have been and are being developed. A popular one of these is the least absolute shrinkage and selection operator (Tibshirani 1996), given by

$$\beta(\lambda) = \arg\min_{\beta}(n^{-1}\| Y - X\beta\|^2 + \lambda\|\beta\|_1). \tag{4.3}$$

In addition to discussing this method theoretically, Buhlmann applies it to finding the most relevant motifs from over 250 candidate sequences for explaining the binding strength of a transcription factor on about 200 regions in a handful of ChIP-chip experiments.

(ii) *Graphical models*

Whenever we have a high-dimensional vector of numerical variables $(X_1, \ldots, X_j)$, we can model these as jointly Gaussian, unrealistic though this may be. If, as usual, we are interested in modelling their interdependencies this leads to models of their covariance or its inverse matrix. However, this can be represented as the weights on the edges of a graph whose vertices are the variables. Thus, $\mathrm{cov}(X_i, X_j)$ is the weight attached to the edge between X_i and X_j. We can, in the case of both numerical and categorical variables, make each variable correspond to a vertex, with the presence of an edge indicating dependence, or for the inverse conditional dependence, though, in general, no single set of edge weights can summarize the degree of dependence. The most striking examples of graphical models in genomics are regulatory pathways (Segal *et al.* 2003, 2008; Meinshausen & Buhlmann 2006; Lee *et al.* 2008).

It has become clear that the graphical dependence structure has to be very sparse for us to be able to estimate it with a small number of replicates (Meinshausen & Buhlmann 2006). Sparsity here means a small number of edges and/or the possibility of reduction in the number of vertices which bear on the question of interest. In fact, it has been demonstrated in the literature that many biochemical and genetic networks are not fully connected (Jeong *et al.* 2001; Gardner *et al.* 2003; Tegner *et al.* 2003). Many genetic interaction networks contain many genes with few interactions and a few genes with many interactions. This observation was made at the workshop, for the protein–protein network in Brunak's talk (Lage *et al.* 2007), the gene network in Marcotte's talk (Lee *et al.* 2008) and the metabolism network in Luscombe's talk (Seshasayee *et al.* 2009). It would appear that genetic networks are intrinsically sparse. A large body of literature deals with questions such as these. On the other hand, regulatory networks function, even when some elements of the pathway have been eliminated, and thus, in some form, exhibit extensive redundancy (Laney & Biggin 1996; Kafri *et al.* 2005). Combining these contradictions is a novel challenge. The book edited by Jordan is an excellent general reference on the construction and fitting of graphical models (Jordan 1999).

A set of issues one has to be cautious about in connection with graphical models is the assignment of causality. This can be viewed as making the graph directed, by assigning arrows to the edges in the model. The arrow, directed from X_i to X_j, means that X_i causes X_j. A thought-provoking discussion of these issues (which are not limited to the high-dimensional context) may be found in Freedman (2005). Of course, conjectured causal arrows can to some extent be validated by additional experiments, e.g. by checking the phenotypic effects of perturbation of possible causal genes (Lee *et al.* 2008), but believing in them purely on the basis of evidence, coming from a body of data designed to measure association, is dangerous.

(iii) *Latent variable models*

This class encompasses a wide variety of models with other names, hidden Markov models (HMMs), mixture models, factor analysis and Bayesian hierarchical models. The common feature of all of these is that the model postulates the existence of unobservable random quantities (latent variables) which, if known, would result in simple stochastic relationships between the observables. The most widely used of these models in bioinformatics is the HMM. A typical formulation is that time corresponds to genomic position and that the corresponding basepair is viewed as an observable. The observable outcomes in a genomic stretch are viewed as conditionally independent, given a latent (hidden) Markov chain of unobservable states evolving in parallel, with the distribution of a basepair at a given position dependent only on what state the Markov chain is in that position. For example, a state of the Markov chain could be the presence of one of several possible transcription-factor binding-sites or absence of any such site. The data can then be used to fit the parameters of the HMM and then predict the most likely state for the Markov chain at a given position. This is part of the model in Segal's talk at the workshop for predicting spatial expression patterns in the *Drosophila* embryo from binding site information, spatial concentration data for several transcription factors, and sequence data for several target genes (Segal *et al.* 2008). HMMs are also an intrinsic part of the SUNFLOWER set of algorithms presented in Birney's talk at the workshop. These algorithms model regions of the genome believed to serve as gene regulatory elements, that is, genomic regions that provide binding sites for transcription factors. These formulations are typical of latent variable models. These models have value for predictive and exploratory purposes. Unfortunately they typically reflect the biology only crudely. For instance, in some biological systems, motif positions have been shown not to be independent (Bulyk *et al.* 2002), and there is no reason to believe that the sequence of binding sites along the genome is Markovian (of any low order). As a result, validatory statements are questionable.

HMMs can be reasonably stably fitted when the number of possible states is small compared to the length of genome considered. Otherwise the 'curse of dimensionality' operates in both stability of fitting and computation since the number of parameters to be fitted and the time needed for fitting algorithms both scale as the square of the number of states (Bellman 1957). HMMs have been used with great success in speech recognition and other engineering and physical science applications since the 1970s (Rabiner 1989). Durbin *et al.* (1999) remains an excellent reference for applications in bioinformatics.

Another well-known example of a hidden variable model is the mixture model in which the hidden variable is the label of a component of the mixture. This model has been extensively used for modelling due to its flexibility and simplicity. By estimating the distribution of individual components and the latent label for each individual, this method provides a useful tool for clustering observations and exploring scientifically meaningful structures in biological problems. For instance, it has been used for clustering subpopulations in human population (Pritchard *et al.* 2000). These are situations where it is plausible to model observed populations as being mixtures of distinct types although the structure of the individual populations being sampled from is less secure. However, again the purpose of the analysis is more exploratory and predictive than validatory. Some general references are available in the literature (Titterington *et al.* 1985; Lindsay 1995). The dangers of overfitting are indirectly present in the issue of the number of clusters to be chosen.

Continuous latent variable models play an important part in dimension reduction. Thus, PCA may be viewed as a model where each of the p observations is of the form $X = AZ$, where A is an orthonormal matrix and Z has a distribution with diagonal covariance matrix. A is the set of population principal components and the variances of Z are the eigenvalues of the covariance matrix of X. The Z's are the latent variables here and dimension reduction corresponds to all but $s \ll p$ of the Z's having variance 0. We can go further if we assume all the components of Z independent and at most one of these Gaussian. Then A can be retrieved up to scaling and the permutation of the rows, using algorithms such as independent component analysis, where again the Z's are latent. A final method of this type is factor analysis, where the observed variables are modelled as linear combinations of Z components plus an error term. If the dimension p is large the usual empirical estimates may be quite misleading. Again, if sparsity is present, procedures that do not suffer from the 'curse of dimensionality' should be used. Sparsity can be enforced either directly through thresholding methods (Bickel & Levina 2008), or through appropriate prior distributions on the above matrices. Although the latter approach is not fully understood theoretically, the success of applications can be judged by predictive performance. For example, studies of cancer genomics are often concerned with predictive/prognostic uses of aggregate patterns in gene expression profiles in clinical contexts, and also the investigation and characterization of heterogeneity of structure related to specific oncogenic pathways. At the workshop, West presented case studies drawn from breast cancer genomics (Carvalho *et al.* 2008). They explore the decomposition of Bayesian sparse factor models into pathway subcomponents, and how these components overlie multiple aspects of known biological structure in this network, using further sparse Bayesian modelling of multivariate regression, ANOVA and latent factor models.

(iv) *Bayesian networks*

A Bayesian network is often just another name for a graphical model that encodes the joint probability distribution for a large set of variables. The term Bayesian applies if the joint distribution of all variables in the model is postulated with at least one, the 'prior' variable, being latent. Its appeal lies in its generality enabling the integration of prior knowledge and diverse data. Its formulation often involves the insertion of causal arrows and, in principle, predicting the consequences of intervention. These models are not mechanistic and reflect biological knowledge only crudely, but they can have predictive value with external validation. Internal claims of causality and evidence have to be taken with a grain of salt. As examples, Bayesian networks have been used to integrate protein–protein interaction data with clinical diagnosis in order to infer the functional groups of proteins (Lage *et al.* 2007). In the work of Huang presented at the workshop, a Bayesian network was built to link published microarray data with clinical databases for medical diagnosis. Recently, a network covering most *Caenorhabditis elegans* genes has been generated using a Bayesian network, which successfully predicted how changes in DNA sequence alter phenotypes (Lee *et al.* 2008). An excellent reference on Bayesian networks is Heckerman (1999). It may well be that, in the instances discussed, most involve a large number of variables. Using dimension reduction methods, as by Buhlmann, might have been beneficial.

As these examples illustrate, probabilistic modelling is used in genomics primarily for prediction. Insofar as the predictions can be verified experimentally, statistical validation is not an issue, though evidently the more biological information a model can successfully mirror the greater its predictive power. Using sparsity leads to good predictive behaviour. However, as discussed below, probabilistic models are essential for any validatory statements.

(*d*) *Validatory statistics*

(i) *Testing for association*

In genomics, the following situation is typical of many recent studies, especially consortium studies, such as ENCODE or genome-wise association studies (Stranger *et al.* 2007; The Wellcome Trust Case Control Consortium 2007; The ENCODE Project Consortium 2007; Barnes *et al.* 2008). Several features, or annotations, are defined across the genome. These features may be putative exons as determined by mRNA-seq or a microarray experiment, transcription-factor binding sites as predicted by a ChIP-seq assay, or perhaps just a measure of local G-C content. The researcher wishes to understand the relationship between two features. This can be stated as a question, such as, 'Do all these new exons predicted by biochemical assays tend to occur in regions predicted to be bound by RNA polymerase?'

In order to answer questions regarding the association of features, one must construct some kind of model for the genome. Once a model has been selected, one can compute confidence intervals, or conduct testing of null hypotheses. It is customary to cite small p-values under the hypothesis of no association as evidence of strength of association, or to compare small p-values in order to argue that one set of associations are stronger than another. There are a number of problems with these practices.

(i) The null model of no association between two features is difficult to formulate mathematically since, for a single genome, association has to be defined taking into account genomic structure.

(ii) In almost all talks at the workshop, associations are being examined and their strength measured for many pairs of factors. To have p-values support many statements of association they have to be very small, since they need to reflect either a Bonferonni correction or figure in an FDR, topics which will be discussed later under the heading of multiple testing (Benjamini 2008; Efron 2008). Not all the papers presented at the workshop were careful on this point.

(iii) p-values are poor measures of association since they measure the distance in some peculiar metric from the implausible hypothesis of no association. Is a p-value of 10^{-12} for an association 1 000 000 times as strong as one with 10^{-6} or only twice as strong (as measured on a logarithmic scale)?

(iv) It is quite unclear how features can be combined in regulatory pathways using p-values.

Pursuing point (i) above, formulations based on the assumption (made in BLAST, in naive phylogenetic models, in position weight matrices and elsewhere) that positions on the genome are independent and identically distributed are evidently unrealistic. Other implicit models such as in homogeneous Poisson processes of features, permitting some basepair clumping, or Markov models are based on convenience rather than a conscious effort to capture underlying structure.

In a paper under submission, presented at the workshop (Bickel's talk), a non-parametric model for the statistical structure of the genome was proposed. This model when used for testing for association between features makes the minimal set of assumptions needed for genuine probabilistic inference based on single genomes. It results in the most conservative assessments of significance of observed association among the models we have cited. Moreover, it provides error bars for measures of strength of association, such as feature overlap. This approach is much more suited to the elucidation of feature interaction and avoids the weaknesses pointed out in points (ii) and (iii) above. As for point (iv), tools, such as graphical and other probabilistic models discussed above, are evidently more suitable. However, the need for an appropriate probabilistic model such as the one in Bickel's talk remains when we are dealing with single genomes.

(ii) *Multiple testing*

In the association analyses discussed above, and more generally during the analysis of large biological datasets, thousands of statistical tests may be conducted, where a number of such tests are expected to be significant. This is the case in, for instance, the analysis of ChIP-seq data: the assay produces a signal that is well defined across the genome. One would like to know where this signal becomes significant, that is, where it deviates from some null distribution (derived analytically or from a negative control). This process is known as 'peak calling', and many solutions have been proposed (Johnson *et al.* 2007; Zhang *et al.* 2008; Rozowsky *et al.* 2009). Many of these rely upon the generation and thresholding of p-values.

In such cases, even if we assume the underlying probability model to be adequate we need to control the possibility of false positives among our statements. This can be handled, in an essentially model-free way, by multiplying the p-value of any test by the number of tests (Bonferroni's inequality). This procedure, which guards against any single false positive occurring, is referred to as controlling the family-wise error rate. It is often seen as much too strict and may lead to many missed findings. An alternative goal is to identify as many significant features in the genome as possible, while incurring a relatively low proportion of false positives.

The FDR of a test is defined as the expected proportion of incorrectly rejected null hypotheses among the declared significant results (Benjamini & Hochberg 1995). Because of this directly useful interpretation, the FDR often provides a more convenient scale than p-values. For example, if we declare a collection of 100 genes with a maximum FDR of 0.10 to be differentially expressed, then we expect around 10 genes to be false positives. This lies between the naive use of single test p-values, and the ultraconservative Bonferroni correction, which controls the possibility of reporting a single false positive in a study. Statistical methods have been proposed either to transform a p-value into an FDR or to compute FDR directly (Storey 2002; Storey & Tibshirani 2003).

Since Benjamini and Hochberg's seminal 1995 paper (Benjamini & Hochberg 1995), several versions of FDR (such as FDR, local fdr, pFDR) have been proposed. These approaches can be combined under a two-component mixture model of true significants and false significants, with the mixture component rate estimated from the empirical distribution of p-values (Benjamini 2008). Thus, a global FDR controls the average number of false positives as above, while a local FDR estimates the posterior null probability for every test. More discussion of their connections and differences can be found in a recent review by Efron and discussants (Benjamini 2008; Efron 2008). In his talk and the paper in this issue (Benjamini *et al.* in press), Benjamini discussed other issues arising in genomic and other datasets using a recently published WTCCC study (Zeggini *et al.* 2007) as an example. These included controlling the error for multiple confidence intervals constructed for selected parameters, measuring replicability in multiple datasets, and strategies to improve the power of testing when the abundance of true findings to be discovered is very low. In particular, he showed how hypotheses may be combined into sets in which they are likely to be true or false together thus increasing power and reducing the number of hypotheses to be tested. The last issue is clearly important and is again a phenomenon of the high-dimensional data we are dealing with.

The FDR is not robust against dependence of hypotheses, precisely the situation which obtains among many genes (Efron 2008). There are attempts to deal with this issue, but they involve using knowledge of the type of dependence which is often not available (Leek & Storey 2008), and more work needs to be done.

(iii) *Methods of inference based on subsampling*

Many methods of inference have a substantial Monte Carlo component, usually labelled as bootstrapping. The bootstrap, introduced by Efron (1979), is a computer-based method for assigning measures of accuracy to statistical estimates. It has become an essential ingredient of many statistical methods. In its most basic form the bootstrap can estimate features of a population, such as quantiles of statistics like the Kolmogorov–Smirnov statistic, which are difficult or impossible to compute analytically. Other applications include the approximation of statistical functions depending on the data, notably including confidence bounds for a parameter. Confidence bounds can be set by estimating the population distribution, either parametrically or non-parametrically, by the empirical distribution of the statistic of interest as computed on each of the bootstrap samples. The general prescription of the bootstrap is to estimate the probability model, and then to act as if it were the truth.

The bootstrap enjoys the advantage, and the danger, of being automatic after the probability mechanism has been estimated. The danger is that it is no better than the hypothesized model. Thus, if we apply it, treating genomic positions as independent and identically distributed, its results can be nonsensical. As a principle, however, it is very important, since it has freed statistics from being unable to deal with situations where there is no closed distributional form. Its justification is always asymptotic (Hall 1997). However, when valid it enables us to deal with situations where the validity of asymptotics is known but the limit is analytically intractable, as in a situation discussed in Bickel's talk at the workshop, testing for uniformity of distribution via a Kolmogorov–Smirnov-like statistic, given that there are many, potentially unknown, genomic regions forbidden to it. The bootstrap has been extended to structured data, where it becomes necessary to simultaneously sample multiple data-units in order to ascertain extant dependencies, such as is done with the model underlying the GSC, the method introduced in Bickel's talk.

(iv) *Bayesian methods*

Bayesian inference is based on posterior distributions. To the models we have discussed, all of which involve unknown parameters, a prior distribution assumed known is added to govern all parameters. We do not enter into the pros and cons of Bayesian inference here. The resurgence of Bayesian methods is due to the possibility of approximately computing posteriors, an analytically infeasible task with most models. This is quite generally done via Markov chain Monte Carlo techniques that characterize the posterior as the stationary distribution of a Markov chain that is run long enough to produce a pseudosample from the posterior. The model-dependent choice of Markov chain and the length of time it needs to be run to approach stationarity are the subject of a great deal of discussion in the Bayesian literature (Gilks *et al.* 1996; Robert 2001).

The problems of high dimension of the parameter space are not resolved by the Bayesian paradigm. Markov chains take much longer to converge to stationarity in high-dimensional state spaces. This is not surprising, since they need to explore the space thoroughly before reaching stationarity. However, it is possible to build sparsity into Bayesian priors producing effective dimension reduction, as for instance in West's talk at the workshop (Carvalho *et al.* 2008). Unfortunately, it is also known that Bayesian methods can behave arbitrarily badly in high-dimensional space (Freedman 2005). The phenomenon that the prior dominates the data can persist for arbitrarily large sample sizes.

The theoretical understanding of Bayesian methods is progressing but unknown pitfalls remain. It is again important to stress that, if Bayesian methods are used for prediction rather than validation, these issues do not arise.

5. Discussion

Our paper has been prompted primarily by the biological panorama presented at the Newton Workshop 'High dimensional statistics and molecular biology'. We have focused (i) on the history of genomics and the technologies developed to study function of the molecular level and (ii) on statistical methodology and its relevance and appropriateness to genomics. In that connection, we have pointed out the following features.

(i) The relative role of explanatory statistics, prediction, probabilistic modelling, and validatory statistics.
(ii) Some dangers of the use of p-values for validation and substitutes for these methods.
(iii) The importance of techniques which assume some sparsity in terms of the relevant variables and of dimension reduction.
(iv) The relatively primitive state of mathematical and statistical modelling in this field.

We have pointed to new methods in statistics, some being currently developed, which address point (iii). Great challenges remain.

We have not discussed modelling at higher levels of organization: intercellular networks, tissues, organisms, populations. The different types of mathematical methods arising naturally in these applications are surveyed sketchily, but extensively, in the report on *Mathematics and 21st Century Biology* from the National Research Council (2005). It is nevertheless clear that the integration of models at the level of the genome with the more mechanistic models of the biology of organisms and the statistical models of population genetics is of great importance and promise.

There is an ever-increasing need for analytical scientists, mathematicians, engineers, physicists and statisticians to enter this exciting and highly interdisciplinary area of computational biology.

We thank Mark D. Biggin for many useful conversations and editing. We thank Nathan P. Boley and Vibhav Viswanathan for assistance in preparation of the article.

References

Alter, O., Brown, P. & Botstein, D. 2003 Generalized singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms. *Proc. Natl Acad. Sci. USA* **100**, 3351–3356. (doi:10.1073/pnas.0530258100)

Altschul, S., Gish, W., Miller, W., Myers, W., Myers, E. & Lipman, D. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1016/S0022-2836(05)80360-2)

Ambroise, C. & McLachlan, G. 2002 Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl Acad. Sci. USA* **99**, 6562–6566. (doi:10.1073/pnas.102102699)

Avery, O., MacLeod, C. & McCarty, M. 1944 Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* **79**, 137–158. (doi:10.1084/jem.79.2.137)

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D. & Hurles, M. 2008 A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.* **40**, 1245–1252. (doi:10.1038/ng.206)

Bartlett, J. M. & Stirling, D. 2003 A short history of the polymerase chain reaction. *Methods Mol. Biol.* **226**, 3–6.

Bellman, R. 1957 *Dynamic programming.* Princeton, NJ: Princeton University Press.

Benjamini, Y. 2008 Comment: microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23**, 23–28. (doi:10.1214/07-STS236B)

Benjamini, Y. & Hochberg, Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* **57**, 289–300.

Benjamini, Y., Heller, R. & Yekutieli, D. 2009 Selective inference in complex research. *Phil. Trans. R. Soc. A* **367**, 4255–4271. (doi:10.1098/rsta.2009.0127)

Bentley, D. 2006 Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16**, 545–552. (doi:10.1016/j.gde.2006.10.009)

Bickel, P. & Levina, E. 2008 Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227. (doi:10.1214/009053607000000758)

Blanchette, M. 2007 Computation and analysis of genomic multi-sequence alignments. *Annu. Rev. Genom. Hum. Genet.* **8**, 193–213. (doi:10.1146/annurev.genom.8.080706.092300)

Blanchette, M. *et al.* 2004 Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715. (doi:10.1101/gr.1933104)

Bor, Y., Swartz, J., Li, Y., Coyle, J. & Rekosh, D. 2006 Northern blot analysis of mRNA from mammlian polyribosomes. *Nat. Protocols.* (doi:10.1038/nprot.2006.216)

Boyle, A., Guinney, J., Crawford, G. & Furey, T. 2008 F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2358. (doi:10.1093/bioinformatics/btn480)

Brudno, M., Do, C., Cooper, G., Kim, M., Davydov, E., Green, E., Sidow, A. & Batzoglou, S. 2003 LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731. (doi:10.1101/gr.926603)

Bulyk, M. 2006 DNA microarray technologies for measuring protein–DNA interactions. *Curr. Opin. Biotechnol.* **17**, 420–430. (doi:10.1016/j.copbio.2006.06.015)

Bulyk, M., Johnson, P. & Church, G. 2002 Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* **30**, 1255–1261. (doi:10.1093/nar/30.5.1255)

Burge, C. & Karlin, S. 1998 Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346–354. (doi:10.1016/S0959-440X(98)80069-9)

Carvalho, C., Lucas, J., Wang, Q., Chang, J., Nevins, J. & West, M. 2008 High-dimensional sparse factor modelling—applications in gene expression genomics. *J. Am. Statist. Assoc.* **103**, 1438–1456. (doi:10.1198/016214508000000869)

Conlon, E. M., Liu, X. S., Lieb, J. D. & Liu, J. S. 2003 Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA* **100**, 3339–3344. (doi:10.1073/pnas.0630591100)

Crick, F. 1958 On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163.

Crick, F. H. 1970 Central dogma of molecular biology. *Nature* **227**, 561–563. (doi:10.1038/227561a0)

Darwin, C. 1900 *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, 6th edn. New York, NY: D. Appleton and Company.

Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. 2002 Capturing chromosome conformation. *Science* **295**, 1306–1311. (doi:10.1126/science.1067799)

Djordjevic, M., Sengupta, A. M. & Shraiman, B. I. 2003 A biophysical approach to transcription factor binding site discovery. *Genome Res.* **13**, 2381–2390. (doi:10.1101/gr.1271603)

Dudoit, S., Gentleman, R. & Quackenbush, J. 2003 Open source tools for microarray analysis. *Biotech. Suppl. Microarrays Cancer: Res. Appl.* 45–51.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. 1999 *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

Efron, B. 1979 Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26. (doi:10.1214/aos/1176344552)

Efron, B. 2008 Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23**, 1–22. (doi:10.1214/07-STS236)

Ewing, B. & Green, P. 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194. (doi:10.1101/gr.8.3.186)

Fisher, R. A. 1930 *The genetical theory of natural selection*. Oxford, UK: Clarendon Press.

Francis, S. C., Michael, M. & Aristides, P. 2003 The human genome project: lessons from large-scale biology. *Science* **300**, 286–286. (doi:10.1126/science.1084564)

Freedman, D. 2005 *Statistical models: theory and practice*. Cambridge, UK: Cambridge University Press.

Gardner, T. S., Dibernardo, D., Lorenz, D. & Collins, J. 2003 Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105. (doi:10.1126/science.1081900)

Gilbert, W. 1978 Why genes in pieces? *Nature* **271**, 491–594. (doi:10.1038/271501a0)

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 *Markov chain Monte Carlo in practice*. London, UK: Chapman & Hall/CRC.

Gilmour, D. & Lis, J. 1987 Protein-DNA cross-linking reveals dramatic variation in RNA polymerase II density on different histone repeats of Drosophila melanogaster. *Mol. Cell. Biol.* **7**, 3341–3344.

Greenbaum, D., Luscombe, N., Jansen, R., Qian, J. & Gerstein, M. 2001 Interrelating different types of genomic data, from proteome to secretome: oming in on function. *Genome Res.* **11**, 1463–1468. (doi:10.1101/gr.207401)

Griffiths, A., Miller, J., Suzuki, D., Lewontin, R. & Gelbart, W. 2000 *An introduction to genetic analysis*, 7th edn. New York, NY: W.H. Freeman.

Haldane, J. 1932 *The causes of evolution*. London, UK: Longman Green.

Hall, N. 2007 Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* **210**, 1518–1525. (doi:10.1242/jeb.001370)

Hall, P. 1997 *The bootstrap and edgeworth expansion*. New York, NY: Springer.

Hartigan, J. A. 1975 *Clustering algorithms*. New York, NY: Wiley.

Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The elements of statistical learning theory*. New York, NY: Springer.

Heckerman, D. 1999 A tutorial on learning with bayesian networks. In *Learning in graphical models* (ed. M. I. Jordan), pp. 301–354. Cambridge, MA: MIT Press.

Ishikawa, D. & Taki, T. 1998 Micro-scale analysis of lipids by far-eastern blot (tlc blot). *Nihon yukagaku kaishi* **47**, 963–970.

Jeong, H., Mason, A. L., Barabasi, S. P. & Oltvai, Z. N. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42. (doi:10.1038/35075138)

Johnson, D., Mortazavi, A., Myers, R. & Wold, B. 2007 Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502. (doi:10.1126/science.1141319)

Jordan, M. I. (ed.) 1999 *Learning in graphical models.* Cambridge, MA: MIT Press.

Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. 2008 Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231. (doi:10.1093/nar/gkn488)

Kafri, R., Bar, A. & Even, Y. P. 2005 Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* **37**, 295–299. (doi:10.1038/ng1523)

Karlin, S. & Altschul, S. 1990 Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* **87**, 2264–2268. (doi:10.1073/pnas.87.6.2264)

Kent, J. 2002 BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664. (doi:10.1101/gr.229202)

Kimura, M. 1983 *The neutral theory of molecular evolution.* Cambridge, UK: Cambridge University Press.

Kulesh, D., Clive, D., Zarlenga, D. & Greene, J. 1987 Identification of interferon-modulated proliferation-related cDNA sequences. *Proc. Natl Acad. Sci. USA* **84**, 8453–8457. (doi:10.1073/pnas.84.23.8453)

Lage, K. *et al.* 2007 A human phenome–interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316. (doi:10.1038/nbt1295)

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. 2004 A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635. (doi:10.1093/bioinformatics/bth294)

Laney, J. D. & Biggin, M. D. 1996 Redundant control of ultrabithorax by zeste involves functional levels of zeste protein binding at the ultrabithorax promoter. *Development* **122**, 2303–2311.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A. G. & Marcotte, E. M. 2008 A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans. Nat. Genet.* **40**, 181–188. (doi:10.1038/ng.2007.70)

Leek, J. & Storey, J. 2008 A general framework for multiple testing dependence. *Proc. Natl Acad. Sci. USA* **105**, 18 718–18 723. (doi:10.1073/pnas.0808709105)

Lindsay, B. 1995 *Mixture models: theory, geometry, and applications.* NSF-CBMS Regional Conference Series in Probablity and Statistics, vol. 5. Hayward, CA: Institute of Mathematical Statistics.

Macinnes, D. 1960 Electrophoresis: theory, methods and applications. *J. Am. Chem. Soc.* **82**, 1519–1520. (doi:10.1021/ja01491a078)

Mardis, E. 2008 The impact of next-generation sequencing technologies on genetics. *Trends Genet.* **24**, 133–141. (doi:10.1016/j.tig.2007.12.007)

McLachlan, G. J., Bean, R. & Ng, S. K. 2008 Clustering of microarray data via mixture models. In *Statistical advances in biomedical sciences: clinical trials, epidemiology, survival analysis, and bioinformatics* (eds A. Biswas, S. Datta, J. Fine & M. Segal), pp. 365–384. Englewood Cliffs, NJ: Wiley.

Meila, M. & Shi, J. 2001 Learning segmentation with random walk. In *Advances in neural information processing systems*, pp. 873–879. Cambridge, MA: MIT Press.

Meinshausen, N. & Buhlmann, P. 2006 Consistent neighbourhood selection for high-dimensional graphs with the lasso. *Ann. Statist.* **34**, 1436–1462.

Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. 1972 Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein. *Nature* **237**, 82–88. (doi:10.1038/237082a0)

Nadler, B., Lafon, S., Coifman, R. R. & Kevrekidis, I. G. 2005 Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In *Advances in neural information processing systems*, vol. 18 (eds Y. Weiss, B. Schölkopf & J. Platt), pp. 955–962. Cambridge, MA: MIT Press.

National Research Council. 2005 *Mathematics and 21st century biology.* Committee on Mathematical Sciences Research for DOE's Computational Biology. Washington, DC: The National Academies Press.

Needleman, S. & Wunsch, C. 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453. (doi:10.1016/0022-2836(70)90057-4)

Ng, A., Jordan, M. & Weiss, Y. 2002 On spectral clustering: analysis and an algorithm. In *Advances in neural information processing systems*, vol. 14 (eds T. Dietterich, S. Becker & Z. Ghahramani), pp. 849–856. Cambridge, MA: MIT Press.

Pritchard, J. K., Stephens, M. & Donnelly, P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rabiner, L. 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286. (doi:10.1109/5.18626)

Robert, C. P. 2001 *The Bayesian choice*. New York, NY: Springer.

Rozowsky, J., Bertone, P., Royce, T., Weissman, S., Snyder, M. & Gerstein, M. 2005 Analysis of genomic tiling microarrays for transcript mapping and the identification of transcription factor binding sites. In *Advances in bioinformatics and computational biology* (eds J. C. Setubal & S. Verjovski-Almeida). Lecture Notes in Computer Science, no. 3594, pp. 28–29. Berlin, Germany: Springer.

Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, Z., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. & Gerstein, M. 2009 PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75. (doi:10.1038/nbt.1518)

Sabo, P. *et al.* 2006 Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods* **3**, 511–518. (doi:10.1038/nmeth890)

Sanger, F., Air, G., Barrell, B., Brown, N., Coulson, A., Fiddes, C., Hutchison, C., Slocombe, P. & Smith, M. 1977 Nucleotide sequence of bacteriophage phi x174 DNA. *Nature* **265**, 687–695. (doi:10.1038/265687a0)

Schena, M., Shalon, D., Davis, R. & Brown, P. 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470. (doi:10.1126/science.270.5235.467)

Schouten, J., McElgunn, C., Waaijer, R., Zwijnenburg, D., Diepvens, F. & Pals, G. 2002 Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* **30**, e57. (doi:10.1093/nar/gnf056)

Segal, E., Shapira, M., Regev, A., Pe'er, D., Bostein, D., Koller, D. & Frienman, N. 2003 Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176. (doi:10.1038/ng1165)

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. 2008 Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540. (doi:10.1038/nature06496)

Seshasayee, A. S., Fraser, G. M., Babu, M. M. & Luscombe, N. M. 2009 Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*. *Genome Res.* **19**, 79–91. (doi:10.1101/gr.079715.108)

Shi, J. & Malik, J. 2000 Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905. (doi:10.1109/34.868688)

Shiraki, T. *et al.* 2003 Cap analysis of gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15 776–15 781. (doi:10.1073/pnas.2136655100)

Smith, T. & Waterman, M. 1981 Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197. (doi:10.1016/0022-2836(81)90087-5)

Smith, L., Sanders, J., Kaiser, R., Hughes, P., Dodd, C., Connell, C., Heiuer, C., Kent, S. & Hood, L. 1986 Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679. (doi:10.1038/321674a0)

Southern, E. M. 1975 Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**, 503–517. (doi:10.1016/S0022-2836(75)80083-0)

Stigler, S. M. 1986 *The history of statistics*. Cambridge, MA: Harvard University Press.

Storey, J. 2002 A direct approach to false discovery rates. *J. R. Statist. Soc. Ser. B* **64**, 479–498. (doi:10.1111/1467-9868.00346)

Storey, J. & Tibshirani, R. 2003 Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445. (doi:10.1073/pnas.1530509100)

Stranger, B. *et al.* 2007 Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853. (doi:10.1126/science.1136678)

Tegner, J., Yeung, M. K., Hasty, J. & Collins, J. J. 2003 Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl Acad. Sci. USA* **100**, 5944–5949. (doi:10.1073/pnas.0933416100)

The ENCODE Project Consortium. 2004 The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306**, 636–640. (doi:10.1126/science.1105136)

The ENCODE Project Consortium. 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816. (doi:10.1038/nature05874)

The Wellcome Trust Case Control Consortium. 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678. (doi:10.1038/nature05911)

Tibshirani, R. 1996 Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–288.

Tindall, K. & Kunkel, T. 1988 Fidelity of DNA synthesis by the thermus aquaticus DNA polymerase. *Biochemistry* **27**, 6008–6013. (doi:10.1021/bi00416a027)

Titterington, D. M., Smith, A. F. M. & Makov, U. E. 1985 *Statistical analysis of finite mixture distributions.* New York, NY: Wiley.

Toth, J. & Biggin, M. 2000 The specificity of protein–DNA crosslinking by formaldehyde: *in vitro* and in *Drosophila* embryos. *Nucleic Acids Res.* **28**, e4. (doi:10.1093/nar/28.2.e4)

Towbin, H., Staehelin, T. & Gordon, J. 1979 Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl Acad. Sci. USA* **76**, 4350–4354. (doi:10.1073/pnas.76.9.4350)

Tukey, J. W. 1962 The future of data analysis. *Ann. Math. Statist.* **33**, 1–67. (doi:10.1214/aoms/1177704711)

Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. & Sidow, A. 2008 Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–835. (doi:10.1038/nmeth.1246)

Velculescu, V., Zhang, L., Vogelstein, B. & Kinzler, K. 1995 Serial analysis of gene expression. *Science* **270**, 484–487. (doi:10.1126/science.270.5235.484)

Venter, C. *et al.* 2001 The sequence of the human genome. *Science* **291**, 1304–1351. (doi:10.1126/science.1058040)

Watson, J. D. & Crick, F. H. C. 1953 A structure for deoxyribose nucleic acid. *Nature* **171**, 737–739. (doi:10.1038/171737a0)

Weiling, F. 1991 Historical study: Johann Gregor Mendel 1822–1884. *Am. J. Med. Genet.* **40**, 1–25. (doi:10.1002/ajmg.1320400103)

Wright, S. 1930 The genetical theory of natural selection: a review. *J. Heredity* **21**, 340–356.

Yang, Y., Buckley, M., Dudoit, S. & Speed, T. 2002 Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Statist.* **11**, 108–136. (doi:10.1198/106186002317375640)

Zhang, Y. *et al.* 2008 Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137. (doi:10.1186/gb-2008-9-9-r137)

Zeggini, E. *et al.* 2007 Viral population estimation using pyrosequencing. *Science* **316**, 1336–1341. (doi:10.1126/science.1142364)

Zerbino, D. R. & Birney, E. 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829. (doi:10.1101/gr.074492.107)