

# An Overview of Recent Developments in Genomics and the Statistical Methods that Bear on Them

BY PETER J. BICKEL<sup>1</sup>, JAMES B. BROWN<sup>2</sup>, HAIYAN HUANG<sup>1</sup>, QUNHUA LI<sup>1</sup>

<sup>1</sup> Statistics, University of California, Berkeley, USA

<sup>2</sup> Applied Science & Technology, University of California, Berkeley, USA

\* All authors contributed equally to this work

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. In hopes of calling fresh perspectives to this field, we point to current analytical challenges, where novel methods, or novel applications of extant methods, are presently needed.

**Keywords:**

## 1. Introduction

The Central Dogma of Molecular Biology, as enunciated by Crick, 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 (1). 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 was imagined in the 1950's. Genes were found to account for at most 2% of the human genome's string of 3 billion base pairs. The remaining "non-coding" portion, initially labeled 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 high throughput sequencing, microarrays and their descendants, in vivo imaging techniques, microscopy, and many others have enabled biologists to begin to analyze 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 study of protein-protein interactions; metabolomics operating on the cellular scale; all following genomics, the study of DNA sequences. These processes are tightly linked and the utility of these labels is unclear, see (2) for an amusing discussion.

However, all of these can be viewed as the beginning of attempts 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 tends 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 genetics and genomics research. The occasion initially prompting this article was the recent workshop on High Dimensional Statistics and Biology held at the Newton Institute in Cambridge, March 31 -April 4, 2008. We will largely use the papers and content presented at the workshop as illustrative examples.

This paper is organized as follows. In section 2, we outline the historical development of genetics and genomics research. In section 3, we introduce the various types of biological technologies, and the data being generated and of the biological questions which are being posed. In section 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 section 5, we discuss possible new directions of biological research and point to where new analysis and tools may be needed.

## 2. A Brief History of Genomics

Charles Darwin published “On the Origin of Species” in 1859, outlining the process of natural selection (3). Contemporary with Darwin’s work, a monk named Gregor Mendel was in the process of ascertaining the first statistical theory of inheritance (4). Mendel’s work was not widely read until the turn of the century, but after its popularization his experiments with pea plants provided a quantitative backbone for Darwin’s observations. The science of genetics and perhaps more generally of modern molecular biology, can be said to have begun when Mendel coined the term, “factors”, to describe the then unseen means of conveyance by which traits, such as the tendency to sprout wrinkly or smooth peas, were transmitted from generation to generation. Mendelian rules, discovered by arduous observation, are now regarded as the basic principles of genetics. 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 and developed the statistical framework of population genetics, and so infused the theory of evolution with genetic explanations and corresponding statistical models (5; 6; 7; 8). Other advances in genetic research around the turn of the century include the discovery that chromosomes contain linearly arranged genes, the basic units of inheritance, and chromosomal crossover, the source of genetic recombination.

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 genetic material (9). In 1953, J. Watson and F. Crick discovered the double helical structure of double-stranded DNA, and the relation between its structure and capacity to store and transmit information (10). These and many other discoveries marked the transition from classical genetics to molecular genetics. In 1958, F. Crick first enunciated the central dogma of molecular biology: DNA codes for RNA, which codes for protein (11). 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. Studies were enhanced significantly by these technologies. In 1972, W. Fiers determined the sequence of a bacterial gene (12). In 1977, F. Sanger first sequenced the complete genomes of a virus and a mitochondrion (13). Other efforts from the Sanger group in the 1970-1980s established protocols for sequencing, genome mapping, data storage, and sequence analyses. 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 analyze mounting stores of data.

During the 1980's and 1990's, the polymerase chain reaction (PCR) (14), automated DNA sequencing, and microarrays solidified genomics as a preeminent discipline within the life sciences. In 1987, on the basis of the Sanger method, Applied Biosystems marketed the first automated sequencing machine. Microarray technology, which can accomplish many genetic tests in parallel, evolved from Southern blotting. In 1987, an early version of gene arrays was first used to profile the expression of a collection of distinct DNA sequences in arrays (15). In 1995, miniaturized microarrays for gene expression profiling were introduced (16).

These modern assays enabled biologists to resolve questions at a scale and depth that was not previously possible. Research topics have included the determination of the entire DNA sequence of organisms, the study of intragenomic phenomena such as interactions between loci and alleles within a genome, the construction of fine-scale genetic maps, and of course, the analysis and integration of 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 (17). 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% of the genome sequenced to 99.99% accuracy. We also note the great contribution of Celera Genomics in accelerating the sequencing of the human genome. Many more genomes have been sequenced in the last decade (17). As of Feb 2009, sequences from the genomes of around 250,000 organisms were publically available (18). Most of the sequenced species were chosen because they are problematic disease-causing agents, or well-studied model organisms or promised to become good models, such as the bacteria *Haemophilus influenzae*, the Yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditis elegans*, or the flower *Arabidopsis thaliana*.

With the completion of sequencing projects for many model organisms, molecular biology has entered the post-genomic era. The focus of research has turned

from the determination of sequence and the genetic units of inheritance to systems biology, an interdisciplinary field exploring molecular networks, which are the basic functional blocks of life. Specific goals include investigating a single level of biological organization, integrating different types of information to advance understanding of whole biological systems, or uncovering how biological systems change over time, on scales as short as pico-seconds, or as vast as the evolution of species. Such investigations are frequently combined with large scale perturbations, including gene-based methods (e.g. RNAi, mis-expression of wild type and mutant genes) and chemical approaches using small molecule libraries.

The ENCODE project, begun in September 2003, is a large scale example of systems biology. It aims to identify all functional elements in the human genome (19). A necessary prerequisite of this project is to define “function” in genomics. Prior to the 1970’s, it was believed that a gene was defined by promoter sequence upstream of a “cystron”, a contiguous, transcribed unit that coded for protein. The discovery of exons and introns by several groups in 1977 demonstrated that this simplistic view was inadequate to capture the function of eukaryotic genomes (20). The ENCODE Consortium reported, amongst other things, that many genes, far more than previously established, generated chimeric transcripts: RNA transcripts including two or more neighboring genes (125). If these chimeric elements turn out to have important biological function, then our notion of a gene may be redefined yet again.

Recent technological developments include Next-Generation DNA sequencers, capable of sequencing billions of base pairs of DNA in each automated run (21). These are game-changing technologies that have already produced more data than any other technology in the history of biology. These new platforms have seen diverse applications since their launch only a few years ago. The Thousand Genome Project will soon release the sequence of 1000 individual human genomes. Cheap and rapid sequencing may revolutionize diagnostic medicine by permitting an unprecedented degree of hyper-differentiation in health-care practices. Indeed, companies such as deCODE and and Perlegen are already bringing individual genetic profiling to the medical domain, which will eventually permit more precise dosage control and superior drug choice.

Rapidly evolving and diversifying fields of biological research, coupled with technological advances, have given rise to needs for novel computational or analytical techniques: algorithms for sequence assembly and alignment; methods for normalizing microarray signals or identifying differential gene expression; approaches for cluster/classification analysis; and statistical tools for systematic or integrative analysis of high dimensional, diverse biological data. Such analyses can involve the reconstruction of dynamic systems from the quantitative properties of their elementary building blocks.

Studies dependent upon the integration of multiple data-types are becoming increasingly prevalent, and are paving the way toward understanding complex biological systems, from embryogenesis in fruit flies (78), to tumor genesis in human cell lines (79). Such studies bring great statistical and computational challenges at different levels. For instance, in large-scale collaborations, such as the ENCODE project, genotypic data is generated in different laboratories and hence may not be directly comparable due to platform and systematic variations (125). This problem can be attacked on two fronts: biologists can standardize methodologies, cell

lines, protocols, and so forth; and analysts can attempt to identify and correct for sources of systematic bias. A wide variety of quantitative scientists (computational biologists, statisticians, mathematicians, computer scientists, engineers, physicists, and biochemists) are working to create, refine, and test computational models to integrate various data-types, but many more are needed and welcome. In section 5, we introduce and discuss the modern statistical techniques in terms of their applications to different biological data in various contexts.

### 3. Basic Technologies

In this section we shall first briefly describe these basic experimental methods or protocols as they are known in the biological literature, the level of precision they are expected to attain, and the types of “noise” (experimental and biological variability) that limit them. We will then, in Table 1, exhibit a partial list of fundamental data types, their date of introduction and questions they initially addressed. As we shall see in this section, the methods underlying these data types have been combined in groups reminiscent of the combinatorial complexity of the genome to generate further higher levels of data. Databases such as NCBI ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) and the Ensembl genome browser ([www.ensembl.org](http://www.ensembl.org)) collect the results of thousands of assays and computational experiments and render them as “features”, annotations defined across the genome in genomic coordinates, for public consumption. 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, of the fundamental interactions of genotype and environment that lead to observable phenotypes and disease.

#### (a) *Gel Electrophoresis*

Gel electrophoresis is a general technique for separating molecules according to their rate of traversal across an electrified gel. Oliver Smithies used starch gels to separate different species of protein from a mixture as early as the 1950s (22). Modern applications are generally concerned with separating protein bound DNA from unbound DNA, or separating DNA fragments by length. In these cases, the gel is a polymer and into it are introduced the target molecules, for instance a collection of DNA fragments of unknown length. Electricity is passed through the gel mixture, and electromotive force drives the molecules toward one side of the gel (which side depends on their charge). The mass to charge ratios of the molecules determines their rate of traversal, and for DNA molecules this is precisely a function of their length. After the molecules have been separated on the gel, they are stained with dye and the gel is imaged to produce a set of bands corresponding to sequences of different length.

One well known example of modern application comes from criminology: it is used to perform the restriction fragment length polymorphism assay (RFLP). In that assay, a restriction enzyme, a molecule that cleaves double stranded DNA (dsDNA) only at some particular recognition sequence, is introduced into solution with purified dsDNA. The enzyme cuts the dsDNA wherever its recognition sequence occurs, and thereby introduces a length distribution on the resulting frag-

ments “unique” to the individual. Gel electrophoresis is used to image this length distribution.

It is possible to determine the lengths of fragments down to a single basepair. This property was employed in a variety of DNA sequencing protocols, as we shall see a later section.

### *(b) Blotting*

The Southern Blot assay was developed by E.M. Southern in the 1970’s. It is an inexpensive and rapid means of determining the presence or absence of a particular DNA sequence in a large pool of unknown fragments (23). It combines gel electrophoresis with the tendency of single stranded DNA to bind with its complement. The unknown pool is rendered into short fragments and separated into several pools of sequences of known length through gel electrophoresis. These are made single stranded and “blotted” onto a membrane. The known pool is radioactively tagged, made single stranded and exposed to the membrane. Sequences which stick to the membrane are present in the unknown pool.

A few years after Southern introduced his method of blotting, a group at Stanford developed an analogous method for the analysis of RNA, which they called, in honor of Southern, the Northern Blot (24). This assay permitted the analysis of gene expression, since it would discern whether a particular gene was being transcribed.

Within a few years, the immunoblot, or Western Blot, followed (25). The Western Blot utilizes blotted antibodies to detect the presence of a particular protein. In the 1990’s a “Far-Eastern Blot” was developed for the analysis of lipids (26).

Although initially used qualitatively, blotting can also be used quantitatively in conjunction with other assays. For instance, in the ChIP assay, described below, the Southern Blot can be used to identify oligonucleotide of unknown DNA sequence, that is, to ascertain the signal generated by the experiment. In such applications, work remains to understand and model the variance of the combined assay. Early versions of microarray, described below, evolved from this technology.

### *(c) In vivo Cross-linking, Immunoprecipitation(IP), and ChIP assay*

Immunoprecipitation is a process originating in 1959, by which a protein of interest is precipitated out of a solution by the addition of an antibody, or a construct containing antibodies. In modern experiments, beads of some sort are coated with antibodies with specific affinity for a single protein of interest, and are then introduced into a solution usually containing many proteins and other materials (such as DNA, RNA, etc.). The beads, once isolated, can be used to purify a particular protein from the solution, or to determine its presence, or for other indirect purposes.

In particular, the method is a key ingredient of in vivo crosslinking assays to measure binding of proteins to DNA in living cells. First developed in the 1980s by David Gilmour and John Lis, in this assay, cultured cells or plant or animal tissues are first treated with a crosslinking agent, such as UV light or formaldehyde, which covalently couples endogenous proteins to the DNA sequences they directly contact in vivo (27). This is necessary because proteins bind and release DNA on the scale of minutes to seconds or faster, so steps must be taken to ‘freeze’

the bound state of the system prior to extracting the DNA from the cell nucleus. Subsequently the crosslinked protein/DNA complexes are removed from the cell and then immunoprecipitation is used to precipitate out of solution segments of DNA bound by a particular protein. The identity of the sequences is then determined, in early experiments for a few gene loci using blotting and more recently for all genome regions using microarray (ChIP-chip) or sequencing (ChIP-seq) techniques, described below.

Noise can enter the assay due to a protein of interest being crosslinked to a DNA region it does not directly contact via an intermediary protein or due to cross reaction of the antibody with another protein. Whole genome amplification and microarrays or direct sequencing invariably introduce noise. However, careful controls and validation experiments can estimate the degree of noise from each of these sources and in the best experiments noise levels are well understood. The binding events occur on scales of around 4-10 base pairs, while the techniques' resolution is at best a fifty or so base pairs, and at worst a kilobase pair (28).

In addition to the ChIP assay, cross-linking has been applied elsewhere. The so-called, 3C, 4C, and 5C assays utilize formaldehyde cross-linking to bind segments of chromatin in close physical proximity to one another (29; 30; 31). This warrants some explanation: Beyond its primary structure, the sequence of base pairs, DNA exists as a double-stranded polymer (dsDNA) of relatively uniform secondary (local) structure; the familiar double-helix. The global, or tertiary structure of dsDNA is more complex, and depends upon many species of proteins, chief among which are histones. Histones come in many types, and form tetramers that act like spools, around which dsDNA wraps. Such a spooled object is known as a nucleosome. The complex composed mostly of dsDNA and histones is known as "chromatin". Nucleosomes are subsequently folded into complex structures themselves. In particular, these protein-bound coils of DNA are themselves coiled, a process known as super-coiling. All this structure results in the massive compaction of the total space required to house a eukaryotic genome. An average human chromosome, for instance, has an extended length of around 10cm, but its super-coiled state provides more than a 10,000-fold reduction in length, resulting in around 10m of chromatin. This also means that many "distal" sequences in the genome, sequences that are far apart when chromosomes are viewed merely as their linear sequences, are frequently brought into close proximity.

The 3C, 4C and 5C assays attempt to extract and sequence these proximal elements. The sequenced results are searched for chimeric subsequences, sequences that do not occur in the genome, but are composed of two or more that do. These are indicative of proximity mediated cross-linking events. The assays differ in the way that the cross-linked products are prepared for sequencing, and these differences result in different scales and resolutions. The 3C assay has fine resolution and requires the researcher to select a particular position in the genome to study. It is now used as a validating assay for the whole-genome-scale 4C, which permits the simultaneous isolation of chimeric subsequences from throughout the genome, or the also massively parallel 5C (32; 33).

Immunoprecipitation has also found a variety of uses. For instance, antibodies have been developed that recognize cytosines (C, in the genetic alphabet) that have been methylated (34). Methylation is of biological interest in eukaryotes because methylated DNA tends not to be transcribed. In all studied tumor cell lines, for in-

stance, many important house-keeping genes, genes whose activity is not specific to any particular tissue or organ, have been silenced via this mechanism (35). Another assay known as bisulfite conversion identifies all unmethylated cytosines.

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. nonspecific binding, misleading cross-linking, sequencing or microarray errors). Since such studies are now being carried out on the scale of whole genomes, methods such as the false discovery rate, see Benjamini (2009) for a discussion, play an important role. Models need to be developed to relate such indirect measurements to biochemical quantifications. In the case of ChIP-chip, for instance, a model that relates signal to the Gibbs free energy of binding (of the protein of interest to DNA), is still on the distant horizon, and will likely require both the application of new statistical methods and technological advances.

#### (d) *Polymerase Chain Reaction (PCR)*

PCR can be thought of as an enormously flexible way of making an arbitrarily large number of copies of shorter segments of DNA, anywhere from tens to a few tens of thousands of basepairs (55). The process is a chemical one, in which an enzyme, a variety of DNA polymerase, makes a complimentary copy of a ssDNA molecule out of a solution of free bases C,G,A,T. The process is akin to the one by which chromosomes in cell nuclei are duplicated during division. Just as in cell duplication, the process can be repeated indefinitely, doubling the number of copies with each replication, producing a geometrically increasing number of copies. PCR plays a key role in producing most types of data that we shall discuss. For instance, (i) PCR is an integral part of modern sequencing, and (ii) it provides an easy and immediate means of testing for the presence of a particular DNA sequence without resorting to the Southern Blot. This process facilitates disease diagnosis, in that the presence of bacterial DNA can be detected by PCR long before it is detectable by other methods.

Amplification errors are rare, and at worst on the order of 1 base pair in 9,000. So in most current applications the statistical issues are minor (55). 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" (MLPA) (57), which is used in many different assays, it may be that there exist subtle differences in the rates of amplification of the sequences and a call for more complex statistical modeling.

#### (e) *DNA Sequencing*

Sequencing is a key technique in molecular biology, and the technique that has given rise to genomics, and much of modern genetics. The sequences determined are the deoxyribonucleic acid (DNA) molecules, which constitute the genetic instruction book for all life. For humans, the total genome of 46 chromosomes (molecules of DNA) consists of a total 3 billion basepairs. Each DNA molecule is a double helix of paired bases, A-T (Adenine to Thymine) and C-G (Cytosine to Guanine), with the bases of one strand in the helix corresponding to the other, as above.

Since the 1970's, a wide variety of technologies targeted at DNA sequencing have been developed resulting in gains of many orders of magnitude in speed and accuracy. The human genome project, for instance, took over 13 years with effort from more than 100 laboratories (41), but today it is possible to resequence the human genome in a single lab in a month or two (38). The basic techniques of sequencing, up to the present, combine biology, chemistry, physics and the mathematical sciences.

First, many copies of segments of the genome to be sequenced are generated, with lengths ranging from tens of thousands to millions of basepairs. The "many copies" are generally obtained either by "whole-genome shotgun sequencing" or by "BAC Amplification", amplification by bacterial artificial chromosome (37). There are many problems with BAC amplification, but conceptually, this protocol involves inserting large segments of circularized DNA into the genome of a population of bacteria (as a chromosome), and allowing these to multiply. This circularized DNA can be extracted from the bacteria and copy number can be precisely controlled. The extracted copies are then broken into smaller segments, currently ranging from a few hundred to several thousand basepairs depending on the technology. The whole genome shotgun technique, on the other hand, breaks the genome into small fragments initially (inserts of around 3kb), which are integrated into bacterial plasmids, and then amplified as in BAC. This technique was developed by Celera Genomics during the sequencing of the human genome. Generally speaking, these segments are "read" using fluorescent tagging and scanning techniques.

At this point statistical issues come in. Sometimes the base pair calls are wrong. Error rates vary drastically, and have different consequences depending on the length of individual reads generated by the sequencing methods. Given the reads, one faces a primary computational difficulty: one generates thousands of reads of various lengths, but doesn't know a priori how these reads are supposed to fit together. Sequencing a new genome requires the assembly of 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. In order to solve this problem, a variety of "mapping" or "assembly" algorithms have been developed with varying degrees of success (37). The initial strategies are generally described as consisting of three key steps: overlap, layout, and consensus. In the overlap step, the algorithm attempts, in a computationally tractable fashion, to find all sets of reads that appear to overlap the same subsequence. In the layout step, the fact that various reads overlap is used to assemble them into a partial ordering. Lastly, since base-calling errors will have occurred during sequencing, the "consensus" of "overlapping" subsequences is determined by multiple alignment and some sort of averaging. Fairly sophisticated mathematical tools have been employed (39; 40), but substantial differences between algorithms remain, and depend on the length of the reads and the scale of the piece of the genome to be sequenced.

The frequency with which reads tend to overlap is related to the concept of sequencing "coverage", which is the average number of times each base pair is sequenced. For instance, in the first data release of the 1000 Genomes project, the human genomes were sequenced with, on average, 2X coverage, which is to say that, for each individual, on average, each basepair occurs in two reads. To assemble a genome, much greater depth is required. The 1000 Genomes Project, of course, is re-sequencing the human genome, and hence reads are simply mapped

back to the reference genome, which is to say that large scale structure is known, and does not have to be inferred directly from the data. Furthermore, for reasons as yet poorly understood, sequencing is not a uniform or homogeneous process of sampling subsequences from a longer stretch (e.g. a chromosome). Rather, there is a complex statistical background (51).

After the initial cost of establishing a sequencing pipeline, cost scales linearly with coverage. Many open questions remain regarding the actuarial problem of assessing the coverage to which a given project should be sequenced in order to enable a particular set of inferences. The 1000 Genomes Project would like to map all, or at least many, single nucleotide polymorphisms (SNPs) that occur in at least 1 in 100 individuals (see [www.1000genomes.org](http://www.1000genomes.org)). Is the 2X coverage thus far completed sufficient for this task? The answer, of course, will remain unknown until further rounds of sequencing provide deeper coverage; when the subsampling of reads will permit a rigorous analysis.

Many of the most computationally and mathematically challenging problems in molecular biology continue to center around the sequencing and re-sequencing of genomes. However, programs such as Ewan Birney's Velvet have helped to push back the boundaries of de novo applications of the short-read technologies. Velvet has been used to correctly assemble contiguous mammalian sequences more than 2kb in length from the 30bp reads produced by the Illumina platform. The SOLiD platform makes fewer errors and provides 50bp reads, and hence it may be that eventually genomes are assembled tens of base pairs at a time.

Beyond issues of genome sequencing, these next generation technologies have provided key means of sensing the results of assays that, at some point, require the identification of DNA sequences. As we mentioned above, IP protocols, such as ChIP or bisulfite conversion, are now being combined with sequencing to directly observe the precipitated chromatin. One of the earliest uses of sequencing in this "sensing" capacity was the Serial Analysis of Gene-Expression, introduced in 1995 (42) to capture the relative frequency of transcription in a high-throughput fashion. This method involved utilizing the naturally occurring enzyme reverse transcriptase to "reverse-transcribe" RNA transcripts of genes in cell nuclei into a DNA copy, known as cDNA. This process substantially predated the SAGE protocol, but the insight of SAGE was to use a subsequence/tag, extracted from a unique position, to distinguish between different transcripts.

An updated version of the SAGE assay is known as CAGE, or Cap Analysis of Gene Expression (43). In this version, the 5' end of the gene-transcript is identified and ultimately sequenced. This is advantageous because genes are transcribed from the 5' direction, and hence this method has elucidated the complexity of transcription start sites: a given gene may have many hundreds, or even thousands, of transcript variants, many of which begin at various locations, often outside of "promoter sequence", the idealized region immediately 5' to a transcribed gene responsible for binding the various proteins that make up the transcription and transcriptional activation machinery necessary for gene expression.

Today, sequencing is rapidly becoming the dominant means of sensing in genomic assays. Since modern sequencing technologies involve multiple rounds of amplification by PCR, as we noted above, it will be important to ascertain the component of variance due to these sequencing protocols. For many purposes, such as ChIP-seq assays (44), it also becomes important to successfully map sequenced

reads back to the genome in order to make statistical inferences. This mapping step is an additional source of variance that has yet to be explored in detail.

*(f) Sequence Alignment*

Sequencing technologies gave rise to the analysis of genomic sequence in general, and the genomic analysis of phylogeny in particular. This was an area of inquiry previously the sole domain of morphological biology. For the first time it was possible to assess relationships between species, individuals, and clades via the direct interrogation of the genetic material. During the 1980-1990's, the central aim in 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 (45). Since sequence homology is generally concluded on the basis of sequence similarity, this aim gave rise to the need for computational methods and algorithms, particularly for the discovery of duplicated sequences in the same genome (such sequences are called "paralogs") and highly similar sequences in the genomes of related species (called "orthologs"). To this day, the identification of homologous sequences between genomes is still the primary method of de novo gene annotation in newly sequenced genomes (46).

Early efforts in sequence comparison were to align amino acid sequences, which are generally short, with at most several thousand residues. Sequence aligners attempt to find sequences that differ at only a few positions, or are identical up to insertions or deletions. The Needleman-Wunsch algorithm, published in 1970 (47), was an application of dynamic programming to this problem, and gave an alignment of two sequences optimal under a particular, user-defined, substitution and insertion/deletion matrix. An alignment returned by this algorithm is known as a "global alignment". By the early 1980s, longer sequences, of both DNA and RNA were under study, and the Smith-Waterman algorithm was published in 1981 (48), which generalized Needleman-Wunsch to find optimal alignments of subsequences within longer molecules. This process is called "local alignment". Both of these time-intensive algorithms were precursors to high-throughput technologies capable of searching through millions of sequences and subsequence in order to find homologous elements. The first truly high-throughput tool was BLAST, (Basic Local Alignment Search Tool), developed at NCBI in 1990 (49). BLAST searches a query sequence against a database, consisting of, for example, several genomes, and attempts to detect all elements in the database homologous to at least a subsequence of the query sequence. The algorithm first identifies short exact matches, and then attempts to extend those matches under an internal metric, allowing for substitutions, deletions, and insertions. These matches are reported to the user, usually ordered according to the probability of finding such matches in random sequence under the Karlin-Altschul statistics (86). In 2002 BLAT, (BLAST-Like Alignment Tool), was introduced by Jim Kent of the UCSC Genome Browser, and is essentially a faster and more sensitive version of the original (50).

The problem of locating homologous sequence was initially treated in an almost purely pragmatic fashion. BLAT and BLAST have penalty matrices that can be changed when aligning different species with different anticipated evolutionary relationships. Both algorithms are incredibly fast, and their code highly optimized, since both were designed to cope with the massive objects that are genomes at a

time when computational resources were severely limited by computer technology. As computing power increased substantially throughout the 1990s and into the new millennium, more detailed mathematical models for sequence alignment were proposed.

Tools such as BLAST have made pair-wise alignment quite fast. Multiple alignment, however, is computationally far more expensive. Under even a simple additive metric, obtaining the optimal multiple alignment for a few dozen sequences of more than a few thousand base pairs remains computationally intractable. Hence, a variety of heuristics have been proposed, most of which fall into a class of methods known as “progressive alignment”. These techniques involve the construction of pair-wise alignments coupled with techniques for aggregating the pair-wise data into a multiple alignment (53).

Genome-wide multiple alignment, an increasingly popular aim, increases the complexity of ordinary this task by several orders of magnitude. Here, the idea is to align all orthologous elements from a number of genomes simultaneously. Some aligners allow some elements to occur multiple times. For instance, when aligning fish to mammal genomes, there are many genes active during development that are in single copy in fish that have been duplicated and subfunctionalized one or more times in mammals. Hence, a single copy fish gene may be aligned to several orthologs. Other aligners enforce a well ordering, allowing each sequence to occur once and only once. Others enforce a well ordering on only one species, called the reference genome, which is ‘decorated’ with orthologs. The underlying strategy for each of these approaches is as above: many pair-wise alignments are conducted, and those pair-wise alignments are aggregated. Aggregation is non-trivial due to the size of eukaryotic genomes and the complexity of repeat-structure, and the particular technique varies among methods of multiple alignment (56). A popular example is the Threaded Block-Set Aligner (TBA) (54), which sorts pair-wise alignments found using BLASTZ (a variant of the BLAST algorithm) into a partial ordering, employing a heuristic algorithm to break cycles whenever they form (heuristics are necessary as the problem is NP-hard). Phylogenetic information is utilized as a guide to combine pair-wise local alignments into the global multiple alignment.

#### (g) *Microarrays*

The general concept of a microarray is as follows: a chip is “printed” with tens of thousands of variants of a particular polymer, such as DNA, RNA, or cDNA. Each variant appears in a tiny dot, where each dot contains many copies of the same sequence. A DNA microarray will contain tens of thousands or millions of single stranded DNA sequences, and short sequences of unknown identity, called the target sequence or the sample, will be washed across the chip. This approach has been used as an alternative to more costly sequencing, although the new high-throughput technologies are rapidly changing the cost-landscape.

In each version of the assay, the probes are designed so that, when bound, they either fluoresce or are detected by some other light based imaging method. This is generally accomplished simply by labeling the target sequence with an imaginable tag, so that the intensity of a particular probe is proportional to the relative quantity of the target (or sample) present, permitting a quantitative interpretation of the assay (16).

The many dozens of applications of microarrays share common computational challenges: after the “wet lab” portion of the assay is conducted, what remains is a high-resolution image of tens of thousands of variously illuminated little dots. The worth of the assay is predicated on image processing, and subsequent statistical analysis.

Noise enters from many sources: non-specific binding of target to probe (false positives); unanticipated inaccessibility of certain probes due to, for instance, steric hindrance (false negatives); unknown scaling, the relationship between total measured luminescence and quantity of the target, is generally non-linear (58; 59; 60).

Microarrays have been a principal focus of interest in the statistics and computer science communities. They provide perfect example of the paradigms of modern statistics: Observations with thousands of dimensions, repeated only a few times under possibly different condition, a sample or more likely a time series, but with enormous sparsity present; most of the fluorescent dots are noise. Drawing conclusions as to what is signal and what is noise has developed a rich literature already, for instance, on multiple testing, false discovery rates and related measures (133).

An important application of DNA microarrays is the “tiling array”, where nearly every sequence in a genome, of at least some particular length, e.g. 30bps, occurs on the array. Tiling arrays are among the most powerful tools in genome-wide investigations. They have been used in many applications, such as transcriptome mapping, ChIP-chip. The Bulyk lab (62) has devised a novel use of DNA microarrays, and is presently using them to identify the DNA binding preferences of transcription factors.

More generally, most of the canonical assays in use today were originally made possible by this technology. For example, DNase I, a naturally occurring enzyme that cleaves naked DNA approximately indiscriminately, has long been used to explore the tertiary structure of chromatin (63). That is, DNase is used to identify regions of the genome more or less sequestered by histones and the complex three dimensional packing that permits, for instance, humans to fit chromosomes with an average length of 10cm into cell nuclei a few hundred nanometers in diameter. In this assay, DNase I is introduced to nuclei, and the resulting chromatin fragments are separated by size using gel electrophoresis. The shorter fragments, representing genomic regions not sequestered by histones, are then identified using a microarray. Of course, like many assays the most modern versions take advantage of high-throughput sequencing technologies to directly identify the fragments.

#### *(h) Quantitative Imaging*

A large proportion of genome sequence codes for the differential expression of proteins and RNAs. This is especially so for complex multicellular organisms, plants and animals, where extensive cis regulatory sequences direct extraordinarily complex patterns of spatial and temporal gene transcription across thousands or even billions of cells. Because these patterns change between neighboring cells, often showing quantitative gradations in level (rather than on/off differences), an accurate record of what plant and animal genomes ultimately encode requires the establishment of new atlases that record expression patterns and morphology in a computationally analyzable form. Advances in labeling, imaging and image analysis

are permitting the development of such datasets, providing a framework for systems modeling of how *cis* regulatory information is read out to its final form (64).

A variety of ways to label RNA and protein expression have been developed. Specific RNAs within a tissue can be detected by hybridizing them with chemically modified nucleic acid probes of complementary sequence in so called *in situ* hybridization experiments. Proteins can be labeled using antibodies or by using a reporter gene, which is attached to the gene of interest and the contiguous genetic unit integrated into the genome (65; 66). Classical reporters, such as *lacZ*, require that the organism be killed and stained (67). However, Green fluorescent protein (GFP) and its other color derivatives can be imaged directly as reporters in living cells, allowing “movies” of the dynamics of protein movement and gene expression across a field of cells to be made (68). However, GFP is not practical in many situations or tissues, because they are too murky to image without first clearing and staining.

Various resolution datasets of gene expression have been generated for a variety of model organisms and systems, including the fruit fly embryo (65; 66), mouse brain (69) and zebra fish (70). Some of these datasets provide only lower resolution images, others provide cellular resolution data that has been processed by image analysis to produce spreadsheet style tables recording the changing expression of many proteins and mRNAs in each cell in an organism over time, along with the changing spatial coordinates of the cells. Computational modeling has allowed the likelihood for different potential regulatory interactions within transcriptional regulatory to be explored, revealing unexpected aspects of the system and confirming others (71; 65; 66).

Presently, a new generation of fluorescing molecules is being refined for use in biology. Quantum dots are semiconductors that, like GFP, fluoresce when exposed to particular spectrums of light (74; 75). However, they can be much, much smaller, and flash in rapid intervals instead of producing a constant glow. It is possible that, in the near future, quantum dots will enable studies of cellular activity with nanometer resolution. This would permit, for instance, the direct, visual mapping of the binding habits of transcription factors in individual cells, at individual genomic loci. Unlike GFP, quantum dots are not part of reporter genes, but rather, are used like an organic dye. Advances in semiconductor design will be necessary to generate sufficiently non-toxic quantum dots for wide-spread use in living organisms, and statistical techniques will need to be applied to interpret the resulting signal with the sub-diffraction-limit resolution.

Other techniques, such as Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM) (76), have already been demonstrated to capture images with 20nm resolution. These next-generation imaging technologies require clustering techniques to leverage multiple flashing signals into centroids, representative of stationary sources. With further refinement, these and other assays may eventually unveil molecular interactions in their native environment, whereby, in order to understand the binding affinity of a protein for DNA, we have but to look and see.

(i) *The Integration of Multiple Techniques*

We have chosen to highlight several of the many canonical technologies presently in use in molecular biology in this section. These and other technologies and techniques come together in pairs and higher-order groupings and sequences 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, blotting, or PCR, depending upon the desired scale and resolution of the assay. Other examples can be seen in Table 1.

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. For instance, we cannot yet “watch” the binding of transcription factors to promoters and enhancers, and we cannot observe, in any sense, their direct impact on subsequent gene transcription. In 2008, Eran Segal and colleagues published, “Predicting expression patterns from regulatory sequence in *Drosophila* segmentation” (77), where they attempted to impute regulatory interactions from biological image data of gene expression, ChIP-chip identified transcription factor binding regions, and statistical models of protein/DNA binding affinity. They developed and fit a thermodynamic model of gene expression that explained observed expression patterns from predicted patterns of transcription factor binding. Their approach constitutes an integration of many techniques from biochemistry and statistics, and is likely indicative of the direction of molecular and systems biology in the next decade.

During the last four decades, the scale and resolution of molecular biology has evolved radically. In the early 1970’s it was a struggle to sequence a single base pair of DNA. Today a human genome can be sequenced in a matter of weeks in a single lab. We can simultaneously identify all binding sites of a given protein in a genome (ChIP-chip/seq). We can watch the expression of genes across living embryos. The integration of these and other data-types promises to provide insights into the living machinery of organisms on the level of individual molecular interactions.

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. When studying the expression and function of a particular gene, one may imagine an ideal world in which 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 we rely upon a variety of technologies that produce a host of data types, each with their own idiosyncrasies and signal to noise ratios. A successful study usually requires close collaboration between biologists and quantitative scientists. Assays have been born of a process of iterative refinement, where wet-lab biology is progressively informed by the challenges of data analysis. The ChIP-seq assay is an exemplification 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 (87) 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, this can mean more sophisticated techniques: formal dimension reduction using principal component analysis (PCA); 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 is as informative as possible under cost constraints. For instance, as we noted in Section 3 biological data tend to be very noisy: in genome sequencing, the issues of coverage and error rates are crucial; for gene-expression microarray assays, the number of biological and technical replicates has to be chosen such 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 below is probabilistic modeling. 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 binding affinity constant in a reaction involving reagents  $A$  and  $B$  in thermodynamics,

$$K_A = \frac{[\text{Fraction Bound}]}{[\text{Free } A][\text{Free } B]} \quad (4.1)$$

This is a basic element of probabilistic models for binding of a transcription factor (protein) to a given oligonucleotide (short sequence of DNA). As discussed briefly in the previous section, transcription factor binding constitutes a primary mechanism of gene transcriptional regulation (88).

It is only after we have a probabilistic model of the data that we can talk about validity analysis. Technically, constructing a model based on exploratory analysis and fitting it on the SAME data makes validity 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 don't know enough about biological phenomena to postulate hard and fast models. But eventually there is going to be new data gathered since reproducibility of results is always essential in science. The danger of postulating a model based on poor prior information is much greater, since all validity 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 is 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 (143). For studies involving modern high-throughput technologies (e.g. microarray-based gene expression assays), 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 data set dimension is well beyond visualization. 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. A method implicitly based on Euclidean metrics and volume elements is modeling data as coming from a mixture of  $k$  Gaussian distributions and using likelihood ratios from fitted components to assign cluster membership. An excellent reference for all the above methods is Hartigan's 1975 book (89).

This is far from the only approach to be lifted from other parts of statistics used for clustering. Again in the Euclidean case 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. This method has the added advantage if it succeeds of giving a lower dimensional representation of the data. This is especially critical when analyzing high dimensional and noisy biological signals.

In biology, metrics other than Euclidean distance are often needed. A canonical example in biology 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. Given a metric, there are many clustering techniques, such as: the classical agglomerative and other hierarchical methods;  $k$  means clustering and other disaggregative methods. Many applications of clustering were discussed in the workshop, including classification of differentiation of stem cell (Bertone), cell types from microscopy data (Huber) and different virus types (Beerenwinkel).

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 (90). If the resulting matrix is positive semi definite, the vectors can be identified with functions in a reproducible kernel Hilbert space (RKHS) 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 (149).

A huge literature on clustering using spectral properties of appropriate matrices has developed, in particular with so called graph clustering. The relations of these methods to natural properties of random walks and diffusions on appropriate spaces have been well explored (95; 92; 93; 91). 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 of course being the training sample. There are many types of classification methods. Among the most popular are neural nets, logistic regression, CART, support vector machines and an old chestnut  $k$  nearest neighbor 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, Tibshirani and Friedman (96).

The basic principle behind all these methods is the same. Given a training sample of feature vectors and known outcomes  $(X_1, Y_1), \dots, (X_n, Y_n)$  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, of course, one only has a sample, often quite small in relation to the dimensions 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  $k$ -nearest-neighbor rule classifies the new  $X$  according to the majority category (value) of  $Y$  among the training set  $X$ 's which are the  $k$  nearest neighbors of  $X$  in an appropriate metric, usually the Euclidean one if all features are real valued. 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-neighbor rule, which predicts perfectly in the training sample, but no matter how large the training sample is, does not approximate the Bayes rule.

An interesting presentation of classification in scoring potential motifs showing the value of dimension reduction is given in Buhlmann's talk (in the workshop) in which he modifies a method for regression analysis between expression levels and motif occurrence frequencies (94). These approaches also indicate the value of data integration in the biological context.

### *(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 often are determined not by a single or just few genes, but by the underlying structure of genetic pathways that involves many genes; in the 1000 genome projects, the DNA sequencing information generated from the newly developed high-throughput sequencing technology is aimed to be cross-computed with the data from previous studies, such as HapMap (138), to produce a more complete and detailed category of human genetic variation. Probabilistic models are an excellent way of organizing our thinking in such situations. As such 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. Thus, Gauss introduced the method of least squares as opposed to the method of least absolute deviations favored by Laplace for computational reasons and the proposed Gaussian distribution because it was the one for which least squares estimates agreed with maximum likelihood (97). 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 correctly used in genomics primarily for predictive purposes.

(i) *Regression Models*

In biology, the dimensionality 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 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,

$$Y = X\beta + \varepsilon \quad (4.2)$$

where  $X$  is an  $m \times p$  matrix of the values of the  $p$  predictive variables associated with each of the  $n$  observations and  $\varepsilon$  is a “noise” vector. It is expected that the column vectors  $X$  will be sparse. For instance, if  $Y$  comes from measures relating to phenotype and  $X$  is the vector of expression scores coming from a microarray, most genes will have no bearing on  $X$ .

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. The simplest of these is the Lasso (100), given by:

$$\beta(\lambda) = \operatorname{argmin}_{\beta} (n^{-1} \|Y - X\beta\|^2 + \lambda \|\beta\|_1) \quad (4.3)$$

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 ( $X_1, \dots, X_j$ ) for instance, expression levels of different genes, we can model these as jointly Gaussian, unrealistic though this may be. If, as usual, we are interested in modeling their interdependencies this leads to models of their covariance structure. But this can be represented as the weights on the edges of a graph whose vertices are the variables. Thus,  $\operatorname{cov}(X_i, X_j)$  is the weight attached to the edge between  $X_i$  and  $X_j$ .

Often, in the case of both numerical and categorical variable, simply the presence of an edge indicating dependence is of greatest initial interest. The most striking examples of graphical models of this type are regulatory pathways (98; 99; 103; 102). Graphical models really correspond to a representation of the joint distribution of the  $X_i$  to be modeled and thus encompass all models. However, thinking in this generality is useful both from a data analytic and computational point of view. The book edited by Jordan is an excellent reference (144).

It has been made clear that the graphical dependence structure has to be very sparse for us to be able to estimate it with the small number of replicates (99). 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 (106; 105; 104). Many genetic interaction networks contain many genes with few interactions and a few genes with many interactions, such as, the protein-protein network in (90), the gene network in (98), and the metabolism network in (108). It would appear that genetic networks are intrinsically sparse and the corresponding precision matrix should be sparse. On the other hand regulatory networks function even when elements of the pathway have been eliminated, and thus in some form exhibit extensive redundancy (110; 109). Combining these contradictions is a novel challenge. A large body of literature deals with questions such as these. Otherwise one runs the risk of fitting so many parameters that inferences may in part be the result of noise.

Another set of issues one has to be cautious about is the assignment of causality, which can be viewed as making the graph directed. That is we assign arrows to the edges with the interpretation that if an edge is directed from  $X_i$  to  $X_j$  then  $X_i$  causes  $X_j$ . A thought provoking discussion of these issues (which are not limited to the high dimensional context) may be found in the Freedman book (112). 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 (98), but believing in them purely on the basis of evidence coming from a body of data designed to measure association is dangerous.

### **Latent Variable Models**

This class encompasses a wide variety of models with other names, Hidden Markov models, 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 relations between the observables. The most widely used of these models in bioinformatics is the Hidden Markov Model (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 with probability of the base pair corresponding to its frequency in a position weight matrix for the motif, or background frequency for absence. 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 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 (102). HMMs are also an intrinsic part of the SUNFLOWER set of algorithms presented in Birney's talk in 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. They model hidden functional states of the genome that are linked through parametric probability models to each other and to the observables. The models are fit using the observables and

the hidden states predicted using the fitted parameters and the probability model. 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 (111), and there is no reason to believe that the sequence of binding sites along the genome is Markovian (of any low order). And so validatory statements are questionable.

HMMs can be reasonably stably fit 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 (113). HMMs have been used with great success in speech recognition and other engineering and physical science applications since the 1970's (114). Durbin (1998) (150) 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 modeling 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 genes with different expression patterns using microarray experiments (145), and for clustering subpopulations in human population (115). 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 literature (116; 117).

Continuous latent variable models play an important part in dimension reduction. Thus, principal component analysis (PCA) may be viewed as a model where each of the  $p_{x_1}$  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 (ICA), where again the  $Z$ 's are latent. A final method of this type is factor analysis, where the observed variables are modeled as linear combinations of  $Z$  components plus an error term. If the dimension  $p$  is large the usual empirical estimates may be quite misleading (118). 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 (119), or through appropriate prior distributions on the above matrices,  $A$ . 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 often are 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 spe-

cific oncogenic pathways. In the workshop, West presented case studies drawn from breast cancer genomics (120). 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 modeling of multivariate regression, ANOVA and latent factor models.

### **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 (90). In Huang’s work presented in the workshop, a Bayesian network was built to link published microarray data with clinical databases for medical diagnosis. Recently, a network covering most *C. elegans* genes has been generated using a Bayesian network, and it successfully predicted how changes in DNA sequence alter phenotypes (98). An excellent reference on Bayesian networks is by Heckerman (146).

In all of these cases the ultimate validation was external so that these can be viewed as use of the models for prediction. It may well be that in all of these instances involving a large number of variables using dimension reduction methods as in Buhlmann might have been beneficial.

As these examples illustrate, probabilistic modeling 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 behavior. 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 (126; 122; 123; 125). 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, for example, “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 null model for randomness on the genome. Once a model

has been selected, one can compute confidence intervals, or conduct testing. 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.

1. The model of no association insofar as it is specified as a probability model is implausible (see below).
2. In any case  $p$ -values are approximations which even for the “correct model” are untrustworthy precisely when they are extremely small.
3. In almost all of the papers at the conference 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 (133; 132). Not all the papers presented in the workshop were careful on this point.
4.  $p$  values are poor measures of association since they measure the distance in some peculiar metric from the implausible hypothesis of no association. Is an association with a  $p$ -value of  $10^{-12}$  1,000,000 times as strong as one with  $10^{-6}$  or only twice as strong (as measured on a logarithmic scale)?
5. It’s quite unclear using  $p$ -values how features can be combined in regulatory pathways.

Point 1 above faces the difficulty that, for a single genome, independence has to be defined taking into account genomic structure. So, formulations which make the assumption made, as, 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 presented at this conference (Bickel’s talk) and to be submitted, several of us with collaborators proposed a non parametric model called the Genome Structure Correction (GSC) for the genome, essentially giving the minimal set of assumptions permitting genuine probabilistic inference based on single genomes. This model permits not only correct (most conservative among models discussed above) assessments, but the other types of inference we discuss below.

In relation to point 3, estimates of association strength such as feature overlap with appropriate error bars are much more suited to the elucidation of feature interaction. As for point 5, for examining several features together there are better tools such as graphical and other probabilistic models discussed above. However, the need for an appropriate probabilistic model such as the GSC remains when we are dealing with single genomes. All such methods have to contend still with the basic problem of such data: The large number of features many of which interact weakly if at all.

(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 analysis: the assay produces a signal that is well-defined across much of 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 (80; 81; 82; 83; 84; 85). 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 Bonferonni’s inequality, by multiplying the  $p$ value of any test by the number of tests. This procedure, which guards against any single false positive occurring, is referred to as controlling the family-wise error rate (FWER). 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 false discovery rate (FDR) of a test is defined as the expected proportion of incorrectly rejected null hypotheses among the declared significant results (129). Because of this directly useful interpretation, the FDR often provides a more convenient scale on which to work 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 ultimately conservative Bonferonni correction, which can be used to control the possibility of discovering a single false positive in a study under the most conservative assumptions. Statistical methods have been proposed either to transform a  $P$ -value into an FDR or to compute FDR directly (130; 131).

Since Benjamini and Hochberg’s seminal 1995 paper (129), several versions of FDR (such as FDR, local fdr,  $p$ FDR) have been proposed. These approaches are similar in the asymptotic sense and they can each be viewed as a two-component mixture model of true significance and false significance, with the mixture component rate estimated from the empirical distribution of  $p$ -values (133). Briefly, a global FDR controls the average number of false positives among the selected, while a local FDR evaluates the posterior null probability for every individual test. More discussions on their connections and differences are in a recent review by Efron and discussants (132; 133).

As it is pointed out, the FDR is not a refuge against dependence, precisely the situation which necessarily obtains among many genes (132). There are attempts to deal with this issue, but they involve using knowledge of the type of dependence which is often not available (136). The issue is clearly important and is again a phenomenon of the high dimensional data we are dealing with.

(iii) *Methods of Inference Based on Subsampling*

All of the above methods have a substantial Monte Carlo component usually labeled as bootstrapping, other than in Bayesian models where the Markov Chain Monte Carlo methods we discuss below are central.

The bootstrap, introduced by Efron, is a computer-based method for assigning measures of accuracy to statistical estimates (137). 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 nonparametrically, 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 are no closed distributional forms. Its justification is always asymptotic (147). 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, 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.

(iv) *Bayesian Methods*

Bayesian inference is based on posterior distributions. To the models we have discussed, all of which involve unknown parameters, is added a prior distribution governing all parameters which is assumed known. 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 (MCMC) techniques that characterize the posterior as the stationary distribution of a Markov chain that is run long enough to produce a pseudo sample 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 (148; 128).

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, see for in-

stance the material presented in West's talk in this conference that we have already discussed (120). Unfortunately, it is also known that Bayesian methods can behave arbitrarily badly in high dimensional spaces (112). 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's 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 set of issues raised at the Newton Conference on "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 its consequences for phenotype at the level of organisms, and (ii) on statistical methodology and its relevance and appropriateness to modern biological contexts.

In that connection, we have pointed out:

1. The relative role of explanatory statistics, prediction, probabilistic modeling, and validatory statistics
2. Some dangers of the use of p-values for validation and substitutes for these methods
3. The importance of techniques which assume some sparsity in terms of the relevant variables and of dimension reduction
4. The relatively primitive state of mathematical and statistical modeling in this field.

We have pointed to new methods in statistics, some being currently developed, which address point 3. Great challenges remain. We have referred repeatedly to the power of integration of different types of data, for the investigation of function at the genomic level as well as prediction of phenotype. This calls for new models and methods. We illustrate from our experience with the data of the Berkeley Drosophila Transcriptional Network Project (see <http://bdtnp.lbl.gov/Fly-Net/>).

The BDTNP combines for a large number of transcription factors for a developmental stage of drosophila among other types of data

1. In vitro protein/DNA binding affinity data
2. In vivo ChIP-chip data
3. Expression data for a number of time points on a cell by cell basis

The ultimate goal is to dynamically describe the interaction of these factors in producing particular developmental.

Essentially, this involves models for "registering" expression of many embryos on an idealized embryo (66), sparse models coupling the different types of data together, sparse differential equation models for the dynamics, and eventually quantitative models for the regulatory networks and their evolution in time. There has, of

course, been a great deal of work on network modeling in the biological (152), physical, computer science and social science literatures (151). However, techniques for fitting such models are, we believe, just beginning to be developed and biological networks pose both familiar issues of sparsity and less familiar ones of “redundancy”. We note also that new models will be called for to deal with the three dimensional structure of the genome, as revealed by new biochemical techniques such as the 4C assay described previously.

We have not discussed modeling 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 National Research Council (141). 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, statisticians, to enter this exciting and highly interdisciplinary area of Computational Biology.

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

## References

- [1] Crick, FHC “On Protein Synthesis”. 1958. Symp. Soc. Exp. Biol. XII, 139-163
- [2] Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. “Interrelating Different Types of Genomic Data, from Proteome to Secretome: ‘Oming in on Function”. 2001. Genome Research, 11: 1463-1468
- [3] Charles Darwin (1900) *The Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life*, Edition 6, D. Appleton and Company.
- [4] Weiling, F (1991). Historical study: Johann Gregor Mendel 1822-1884. *American Journal of Medical Genetics* 40 (1): 1-25.
- [5] FISHER, R. A. (1930) *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
- [6] Wright, S (1930) *The Genetical Theory of Natural Selection: a review*. *J. Hered.* 21:340-356.
- [7] Haldane, J.B.S. (1932) *The Causes of Evolution*. Longman Green, London.
- [8] Griffiths, AJF, Miller, JH, Suzuki, DT, Lewontin, RC and Gelbart, WM (2000) *An Introduction to Genetic Analysis*, W. H. Freeman; 7th edition.
- [9] Avery OT, MacLeod CM, and 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. *Journal of Experimental Medicine* 79 (1): 137-158.

- [10] Watson, J. D., and Crick, F. H. C. (1953) A structure for deoxyribose nucleic acid. *Nature* 171:173.
- [11] Crick, F.H.C (1970) Central Dogma of Molecular Biology. *Nature*, vol. 227, pp. 561-563. 7-695.
- [12] Min Jou W, Haegeman G, Ysebaert M, Fiers W (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* 237 (5350): 82-88.
- [13] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265 (5596): 687-695.
- [14] Bartlett and Stirling (2003). A Short History of the Polymerase Chain Reaction. *Methods Mol Biol.* 226:3-6
- [15] Kulesh DA, Clive DR, Zarlenga DS, Greene JJ (1987). "Identification of interferon-modulated proliferation-related cDNA sequences". *Proc Natl Acad Sci USA* 84: 8453-8457.
- [16] Schena M, Shalon D, Davis RW, Brown PO (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". *Science* 270: 467-470.
- [17] Francis S. Collins, Michael Morgan, Aristides Patrinos (2003) The Human Genome Project: Lessons from Large-Scale Biology. *Science* 300, 286 (2003)
- [18] <http://www.ncbi.nlm.nih.gov>
- [19] The ENCODE Project Consortium. "The ENCODE (ENCyclopedia Of DNA Elements) Project". 2004. *Science*. 306(5696): 636-640.
- [20] Gilbert W. Why genes in pieces? 1978. *Nature*. 271(5645):491-594
- [21] Mardis ER. "The impact of next-generation sequencing technologies on genetics". 2008. *Trends Genet.* 24(3):133-141.
- [22] Macinnes DA. Electrophoresis: Theory, methods and applications". 1960. *Journal of the American Chemical Society.* 82 (6): 1519-1520
- [23] Southern, E.M. (1975): "Detection of specific sequences among DNA fragments separated by gel electrophoresis", *J Mol Biol.*, 98:503-517.
- [24] Bor YC, Swartz J, Li Y, Coyle J, Rekosh D. "Northern Blot analysis of mRNA from mammalian polyribosomes". 2006. *Nature Protocols.* 10.1038/nprot.2006.216
- [25] Towbin H, Staehelin T, Gordon J. "Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications". *Proc Natl Acad Sci.* 76 (9) 4350-4354.
- [26] Ishikawa D, Taki T. "Micro-scale analysis of lipids by far-eastern blot (TLC blot)". *Nihon yukagaku kaishi.* 47 (10); 963-970.

- [27] Gilmour DS, Lis JT. "Protein-DNA cross-linking reveals dramatic variation in RNA polymerase II density on different histone repeats of *Drosophila melanogaster*". 1987. *Mol Cell Biol.* 7(9):3341-3344.
- [28] Toth J, Biggin MD: The specificity of protein-DNA crosslinking by formaldehyde: in vitro and in drosophila embryos. *Nucleic acids research* 2000, 28(2):e4.
- [29] Hagge H, Klous P, Braem C, Splinter E, Dekker J, Cathala G, de Laat W, Forn T (2007). "Quantitative analysis of chromosome conformation capture assays (3C-qPCR)". *Nat. Protoc.* 2 (7): 1722-1733
- [30] Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)". *Nat. Genet.* 38 (11): 1348-1354
- [31] Dostie J, Dekker J (2007). "Mapping networks of physical interactions between genomic elements using 5C technology". *Nat. Protoc.* 2 (4): 988-1002
- [32] Simonis M, Kooren J, and de Laat W (2007). "An evaluation of 3C-based methods to capture DNA interactions". *Nat. Methods.* 4 (11): 895-901
- [33] Dekker J, Rippe K, Dekker M, Kleckner N (2002). "Capturing chromosome conformation". *Science* 295 (5558): 1306-1311
- [34] Weber M et al. "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells". *Nat. Genet.* 37: 853-862.
- [35] Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, Andrews TD, Howe KL, Otto T, Olek A, Fischer J, Gut IG, Berlin K, Beck S. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.* 2004;2:e405.
- [36] Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet.* 2006 Dec;38(12):1359-60.
- [37] Venter CJ, et al. "The sequence of the human genome". 2001. *Science.* 291 (5507): 1304-1351.
- [38] Bentley DR. "Whole-genome re-sequencing". 2006. *Current Opinions in Genetics & Development.* 16(6) 545-552.
- [39] Daniel R. Zerbino and Ewan Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008. 18: 821-829
- [40] Pevzner PA, Tang H, Waterman MS, An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* 2001 Aug 14; 98(17):9748-53.

- [41] Francis S. Collins, Michael Morgan, Aristides Patrinos (2003) The Human Genome Project: Lessons from Large-Scale Biology. *Science* 300, 286 (2003)
- [42] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. "Serial analysis of gene-expression". 1995. *Science*. 270(5235) 484-487.
- [43] Shiraki T, Kondo S, Katayama S, et al. "Cap analysis of gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." 2003. *PNAS*. 100(26):15776-15781.
- [44] Johnson DS, Mortazavi A, Myers RM, Wold B. "Genome-wide mapping of in vivo protein-DNA interactions". 2007. *Science*. 316 (5830) 1497-1502
- [45] Kimura M. "The Neutral Theory of Molecular Evolution". 1983. Cambridge University Press. Cambridge.
- [46] Burge CB, Karlin S. "Finding the genes in genomic DNA". 1998. *Curr Opin Struct. biol.* 8: 346-354
- [47] Needleman SD, Wunsch CD. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". 1970. *J Mol biol.* 48 (3): 443-453
- [48] Smith TF, Waterman MS. "Identification of common molecular subsequences". 1981. *Journal of Molecular Biology.* 147: 195-197.
- [49] Altschul SF, Gish W, Miller W, Myers W, Myers EW, Lipman DF. "Basic local alignment search tool". 1990. *Journal of Molecular Biology.* 215(3): 403-410
- [50] Kent JW. "BLAT – the BLAST-like alignment tool". 2002. *Genome Research.* 12 (4): 656-664
- [51] Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls". 2009. *Nature Biotechnology.* 27: 66-75
- [52] Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB. "Benchmarking tools for the alignment of functional noncoding DNA". 2004. *BMC Bioinformatics.* 5:6
- [53] Brudno M, Do C, Cooper G, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. "LAGAN and Multi-LAGAN efficient tools for large-scale multiple alignment of genomic DNA". 2003. *Genome Research.* 13(4):721-731
- [54] Blanchette M, Kent JW, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. "Aligning multiple genomic sequences with the threaded blockset aligner". 2004. *Genome Research.* 14: 708-715
- [55] Tindall KR and Kunkel TA (1988). "Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase". *Biochemistry* 27: 6008-6013

- [56] Blanchette M. "Computation and analysis of genomic multi-sequence alignments". 2007. *Annual Review of Genomics and Human Genetics*. 8: 193-213
- [57] Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. "Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification". *Nucleic Acids Res.* 30 (12): e57
- [58] Dudoit S, Gentleman RC, Quackenbush J. "Open source tools for microarray analysis". *Biotechniques Supplements, Microarrays and Cancer: Research and Applications*. 45-51.
- [59] Yang YH, Buckley MJ, Dudoit S, Speed TP. "Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*. 11(1):108-136.
- [60] Rozowsky J, Bertone P, Royce T, Weissman S, Snyder M, Gerstein M. "Analysis of genomic tiling microarrays for transcript mapping and the identification of transcription factor binding sites". 2005. *Lecture Notes in Computer Science*. Vol. 3594. pp 28-29. Springer. Berlin/Heidelberg.
- [61] Liu XS. "Getting started in tiling microarray analysis". 2007. *PLOS Computational Biology*. 3(10):e183
- [62] Bulyk ML. "DNA microarray technologies for measuring protein-DNA interactions". 2006. *Current Opinions in Biotechnology*. 17(4)
- [63] Sabo PJ, et al. "Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays". 2006. *Nature Methods*. 3:511-518
- [64] Ptashne M, Gann A: *Genes and Signals*. New York: Cold Spring Harbour Press; 2002.
- [65] Fowlkes CC, Hendriks CL, Keranen SV, Weber GH, Rubel O, Huang MY, Chatoor S, DePace AH,
- [66] Simirenko L, Henriquez C et al: A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell* 2008, 133(2):364-374.
- [67] Spear BT, Longley T, Moulder S, Wang SL, Peterson ML. "A sensitive lacZ-based expression vector for analyzing transcriptional control elements in eukaryotic cells". 1995. *DNA and Cell Biology*. 14(7):635-642
- [68] Shav-Tal Y, Darzacq X, Shenoy SM, Fusco D, Janicki SM, Spector DL, Singer RH. "Dynamics of single mRNPs in nuclei of living cells". 2004. *Science*. 304 (5678): 1797-1800.
- [69] Lein ES, et al. "Genome-wide atlas of gene expression in the adult mouse brain". 2007. *Nature* 445: 168-176
- [70] Mueller T, Wullimann MF. "Atlas of early zebrafish brain development". 2005. Elsevier.

- [71] Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, Sharp D, Reinitz J: Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene. 2006. *Nat Genet.* 38(10):1159-1165.
- [72] Fowlkes CC, Hendriks CL, Keranen SV, Weber GH, Rubel O, Huang MY, Chatoor S, DePace AH,
- [73] Simirenko L, Henriquez C et al: A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell* 2008, 133(2):364-374.
- [74] Moerner WE: New directions in single-molecule imaging and analysis. *Proc Natl Acad Sci USA* 2007, 104:12596-12602.
- [75] Cang H, Xu CS, Motiel D, Yang H. "Guiding a confocal microscope by a single fluorescent nanoparticle". 2007. *Opt Lett.* 32(18) 2729-2731.
- [76] Rust MJ, Bates M, Zhuang XW. "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). 2006. *Nat Methods.* 3(10): 793-795.
- [77] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* 2008, 451(7178):535-540.
- [78] Biggin MD, Tjian R: Transcriptional regulation in *Drosophila*: the post-genome challenge. *Funct Integr Genomics* 2001, 1(4):223-234.
- [79] Bergers G, Benjamin LE. "Angiogenesis: Tumorigenesis and the angiogenic switch". 2003. *Nature Reviews Cancer.* 3:401-410.
- [80] Zhang et al. "Model-based Analysis of ChIP-Seq (MACS)". *Genome Biol* (2008) vol. 9 (9) pp. R137
- [81] Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data". *Nat Methods.* 2008 Sep; 5(9):829-35.
- [82] Johnson DS, Mortazavi A, Myers RM, Wold B. "Genome-wide mapping of in vivo protein-DNA interactions". *Science*, 2007 Jun 8;316(5830):1497-502. Epub 2007 May 31
- [83] Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data". *Nucleic Acids Res.* 2008 Sep;36(16):5221-31. Epub 2008 Aug 6.
- [84] Boyle AP, Guinney J, Crawford GE, Furey TS. "F-Seq: a feature density estimator for high-throughput sequence tags". *Bioinformatics.* 2008 Nov 1;24(21):2537-8. Epub 2008 Sep 10.
- [85] Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls". *Nat Biotechnol.* 2009 Jan;27(1):66-75. Epub 2009 Jan 4.

- [86] Karlin S, Altschul SF. "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes". 1990. PNAS. 87: 2264-2268.
- [87] Tukey, J.W. (1962) The Future of Data Analysis. The Annals of Mathematical Statistics, Vol. 33, No. 1 (1962), pp. 1-67
- [88] Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. (2003) A biophysical approach to transcription factor binding site discovery, Genome Res. 13 (2003) (11), pp. 2381-2390.
- [89] Hartigan, J.A. (1975) Clustering algorithms, Wiley, New York.
- [90] Lage K, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. 2007. Nature Biotechnology. 25:309-316
- [91] Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G. (2005) Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators, Neural Information Processing Systems (NIPS), Vol 18, 2005.
- [92] Meila, M. and Shi, J. (2001) Learning Segmentation with Random Walk", Neural Information Processing Systems, NIPS, 2001
- [93] Ng, A., Jordan, M., and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press.
- [94] Conlon, E.M., Liu, X.S., Lieb, J.D., Liu, J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. PNAS. 2003 Mar 18;100(6):3339-44.
- [95] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-905.
- [96] Hastie. T., Tibshirani, R., Friedman, J. (2009) The elements of statistical learning theory, Springer, New York
- [97] Stigler, S.M. (1986) The History of Statistics Harvard University Press
- [98] 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* Nature Genetics 2008 40(2):181-8
- [99] Meinshausen, N. and Buhlmann, P. (2006). Consistent neighbourhood selection for high-dimensional graphs with the lasso. Annals of Statistics.
- [100] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288.
- [101] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., (2004) Least angle regression, Ann. Statist. Volume 32, Number 2, 407-499.

- [102] Segal E, Sadka T, Schroeder M, Unnerstall U, Gaul U, (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*
- [103] Segal, E., Shapira, M., Regev, A., Pe'er, D., Bostein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* 34, 166C176.
- [104] Gardner, T.S., DiBernardo, D., Lorenz, D. and Collins, JJ (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301, 102C105.
- [105] Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41C42.
- [106] Tegner, J., Yeung, M. K., Hasty, J., and Collins, J.J. (2003). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Science of the United States of America* 100, 5944C5949.
- [107] Fraser AG, Marcotte EM (2004) A probabilistic view of gene function. *Nature Genetics*, 36(6):559-64 (2004).
- [108] Seshasayee A.S., Fraser G.M., Babu M.M., Luscombe N.M. (2008) Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*. *Genome Research* 2008 Oct 3.
- [109] Kafri R., A.Bar Even, Y.Pilpel (2005) Transcription control reprogramming in genetic backup circuits *Nature Genetics* 37 295-299
- [110] 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(7): 2303–2311.
- [111] Bulyk ML, Johnson PLF, Church GM. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research* 2002 Mar 1; 30(5):1255-1261.
- [112] Freedman D. (2005) *Statistical Models: Theory and Practice*, Cambridge University Press, 2005
- [113] Bellman RE. (1957) *Dynamic Programming*. 1957. Princeton University Press. Princeton, NJ.
- [114] Rabiner LR. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. 1989. *Proc. IEEE*. 77(2)257-286.
- [115] Inference of population structure using multilocus genotype data. J.K. Pritchard, M. Stephens and P. J. Donnelly, 2000. *Genetics* 155: 945-959.
- [116] Lindsay BG. (1995) *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS Conferences in Probability and Statistics

- [117] Titterton D.M., Smith A.F.M., and Makov U.E. (1985). *Statistical Analysis of Finite Mixture Distribution*. Wiley, Chichester, New York
- [118] Johnstone IM. (2009) High dimensional statistical inference and random matrices. To appear in the Proceedings of the ICM.
- [119] P.J. Bickel and E. Levina (2008). Regularized Estimation of Large Covariance Matrices. *Annals of Statistics* 36(1):199-227.
- [120] C.M. Carvalho, J.E. Lucas, Q. Wang, J. Chang, J.R. Nevins and M. West. "High-dimensional sparse factor modelling - Applications in gene expression genomics." *Journal of the American Statistical Association* 103 (2008):1438C1456.
- [121] M. Belkin and P. Niyogi, (2003) Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15 (6):1373-1396, June 2003
- [122] Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME, A robust statistical method for case-control association testing with copy number variation. *Nat Genet.* 2008;. PMID: 18776912 DOI: 10.1038/ng.206
- [123] Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavar S, Deloukas P, Hurles ME, Dermitzakis ET, Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science.* 2007;315:848-53
- [124] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME, Global variation in copy number in the human genome. *Nature.* 2006;444:444-54.
- [125] The ENCODE Project Consortium, (2007) Identification and analysis of functional elements in 1
- [126] WTCCC, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447 (2007) 661-678. The Wellcome Trust Case Control Consortium.
- [127] Donoho, I. Johnstone, G. Kerkyachrian, and D. Picard, Wavelet shrinkage: Asymptopia?, *J. Roy. Stat. Soc.*, 57, pp. 301–369, 1995.
- [128] Robert, C.P. (2001) *The Bayesian choice*, Springer
- [129] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* 57 289-300.

- [130] Storey JD. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64: 479-498.
- [131] Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100: 9440-9445
- [132] Efron, B. (2008) Microarrays, Empirical Bayes and the Two-Groups Model, *Statistical Science*, 2008, Vol. 23, No. 1, 1C22
- [133] Benjamini, Y (2008) Comment: Microarrays, Empirical Bayes and the Two-groups model, *Statistical Science*, Vol. 23, No. 1, 23-28
- [134] Morris, C (2008) Comment: Microarrays, Empirical Bayes and the Two-groups model, *Statistical Science*, Vol. 23, No. 1, 34-40
- [135] Cai, T. (2008) Comment: Microarrays, Empirical Bayes and the Two-groups model, *Statistical Science*, Vol. 23, No. 1, 29-33
- [136] Leek, J.T. and Storey, JD. (2008) A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105: 18718-18723
- [137] Efron, B.(1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7 (1): 1C26.
- [138] The international HapMap project. <http://www.hapmap.org/>
- [139] Politis, D. N., Romano, J. P., Wolf, M. (1998) *Subsampling*, Springer, New York
- [140] Donnelly, P (2008) Progress and challenges in genome-wide association studies in humans. *Nature Insight* 456 (2008) 728-731.
- [141] National Research Council. *Mathematics and 21st Century Biology*. Committee on Mathematical Sciences Research for DOE's Computational Biology.
- [142] Kim, W.K., Krumpelman, C., Marcotte, E.M. (2008) Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biology*, 2008, 9:S5.
- [143] Ambrose, C & McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray geneexpression data. *PNAS*, 99(10): 65626566.
- [144] Jordan, M.I. (1999) *Learning in graphical models*. The MIT Press, Cambridge, MA.
- [145] McLachlan, G.J., Bean, R., and Ng, S.K. (2008). Clustering of microarray data via mixture models. In *Statistical Advances in Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*, A. Biswas, S. Datta, J.P. Fine, and M.R. Segal (Eds.). Hoboken, New Jersey: Wiley, pp. 365-384.
- [146] D. Heckerman. A Tutorial on Learning with Bayesian Networks. In *Learning in Graphical Models*, M. Jordan, ed.. MIT Press, Cambridge, MA, 1999

- [147] Hall, P. (1997) *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.
- [148] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996
- [149] Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M. I. and Noble, W.S. (2004) A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626-2635
- [150] Durbin R, Eddy S, Krogh A, Mitchison. "Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids". 1999. Cambridge University Press.
- [151] Newman MEJ. "The Structure and Function of Complex Networks". 2003. *Siam Rev.* 45 (167)
- [152] Alon U. "An introduction to systems Biology: design principles of biological circuits". 2006. Chapman & Hall.

Table 1. Illustrative examples of current basic questions of genomics and some of the data used to answer them

Basic questions	Data type	Enabling technology		Typical contemporary scale(s) of data
How do biological sequences differ between individuals, populations, and species?	SNPs (single nucleotide polymorphisms)	Microarray (SNP array)		genome wide
		Sequencing		
	CNV (copy number variation)	Microarray (CGH array)		
		Sequencing		
Evolutionarily conservation	Sequence alignment		one element vs. a database	
When, where, and with what frequency are genes transcribed?	mRNA quantization	SAGE	Sequencing	tens of thousands of loci
		CAGE	Sequencing	
		Sequencing (mRNA-seq)		genome wide
		Microarray (Tiling array)		
	mRNA localization	Imaging		one or several loci
How is gene transcription regulated?	Nucleosome depletion	DNase I	Microarray	genome wide
			Sequencing	
	Histone modification	ChIP	Microarray (ChIP-chip)	
			Sequencing (ChIP-seq)	
	DNA methylation	Bisulfite conversion	PCR (methyl-PCR)	one to several loci
			Sequencing (methyl-seq)	genome wide
	Chromatin tertiary structure	3C		tens or hundreds of loci
		4C		genome wide
5C		hundreds or thousands of loci		
Protein (e.g. transcription factor) binding sites	ChIP	Microarray (ChIP-chip)	genome wide	
		Sequencing (ChIP-seq)		