

Integrative Genomics and Functional Explanation

Jo Davies¹, Thorunn Rafner², Garrett Hellenthal¹ and Jotun Hein¹

¹ Department of Statistics, University of Oxford

² deCODE, Iceland

February 11, 2009

Abstract

In this paper we present a conceptual framework common to all studies of global biological variation. It comprises of three components; data, concepts, and analyses. Most studies are still founded on the collection of one or two data types, but with the recent emergence and deployment of affordable high-throughput biological technologies, an increasing number of studies are reporting integrative approaches. We evaluate the contributions of these studies to biological understanding and conclude with discussion about integrative functional studies as a necessary follow-up to global discovery driven approaches.

Contents

1	Introduction	3
2	Data Types	5
2.1	Genomic Data	5
2.2	Epigenomic Data	6
2.3	Transcriptomic Data	6
2.4	Proteomic Data	8
2.5	Metabolomic Data	9
2.6	Phenomic Data	10
2.7	Other Data Sources	11
3	Concepts	13
3.1	A Mapping from Genome G to Phenome F	13
3.2	Networks	15
3.2.1	Biological Networks	15
3.2.2	Theoretical Network Models	17
3.2.3	Statistical Network Models	18
3.2.4	Biological Interpretation of Networks Inferred Statistically	20
3.3	Genealogical Relationships	21
3.3.1	Genealogies relating cells in an individual	21
3.3.2	Genealogies relating individuals of a population	22
3.3.3	Genealogies relating species	22
3.4	Knowledge	23
3.5	Hidden Structures	24
4	Analyses	26
4.1	Analysis of single sources of data	26
4.1.1	Species Level Genomic Variation Data; (G)	26
4.1.2	Human Genetic Variation Data; (G)	27
4.1.3	Molecular Quantities; (T), (M), (P)	28
4.2	Analysis of phenotype with another source of data	29
4.2.1	Analysis of phenotype with genetic data; (G+F)	29
4.2.2	Analysis of phenotype with molecular data; (F + T), (F + P), (F + M)	30
4.2.3	Analysis of genetic data with molecular data; (G + T), (G + P), (G + M)	33
4.3	Analysis with multiple molecular data types; (T + P), (T + M), (M + P), (T + P + M)	34
4.4	Integrated analysis of phenotype with at least two other sources of data	35
4.4.1	Comparing genetic associations with different phenotypes	35
4.4.2	Integrated Networks	36
4.5	Analysis of all data types across multiple species.	37
5	Functional Explanation	39
5.1	Identifying Causal Genetic Variants	39
5.2	Identifying Causal Pathways and Networks	40
5.3	Forwards and Reverse Genetics	41
6	Conclusion	42

1 Introduction

Studies of complex phenotype and many other studies within biosciences can be decomposed naturally into three components (S1-S3). While it is not a perfect decomposition, it describes the set up surrounding many biological investigations. This review presents, discusses and evaluates the contribution of these components to biological understanding.

S1 **Data:** observations of a biological system.

S2 **Concepts:** provide the foundation for appropriate modelling strategies and interpretation of data.

S3 **Analyses:** provide the formal structure of the modelled system and the statistical framework in which models are fitted to data.

The present revolution in the biosciences is driven by the development and the ongoing improvement of a series of high throughput technologies which can now capture a range of biological information. This provides a rich source of data and paves the way for an unprecedented understanding of organisms and the pathway to genotype and phenotype. Here, we consider six such sources of data (D1-D6) which are described in detail in section 2:

D1 The Genome (G) is the simplest data type, the cheapest, the first and can be measured at the highest accuracy.

D2 The Epigenome (E) comprises the features outside of DNA sequence that affect cellular processes such as structural DNA changes.

D3 The Transcriptome (T) is the level of different transcripts of all genes.

D4 The Proteome (P) is the concentration of proteins and modified proteins.

D5 The Metabolome (M) is the concentration of metabolites.

D6 The Phenome (F) is the set of phenotypic characteristics which comprehensively characterise phenotype of an individual.

The analyses of these data types is founded on one or more of five concepts (C1-C5) we have identified. They are accepted and used so frequently that they are often taken for granted and used without question. First is the assumption that a mapping from genotype to phenotype exists. This motivates many complex phenotype studies and inherent in many such analyses are assumptions about the nature of this function. Second is the general concept of a network. They are used in a wide range of disciplines but in the biosciences, they are used to describe relationships and interactions of biomolecules. Third are genealogies. They can be considered at three levels and are important because they impose structure on genetic material. At the finest level, there are genealogies relating cells within an individual, at the next level there are genealogies which relate individuals of a population, and finally there are genealogies which relate species. Fourth is the concept of knowledge, which is difficult to define and represent but crucial since all studies are founded on it to some degree. At a foundational level, knowledge shapes study design and analytical strategies. At a more specific level, knowledge can be informative for either the construction of specific hypotheses or validation of findings. The most basic example is the Central Dogma of Molecular Biology which describes the flow of information from genotype through to protein (figure 1). While there are accepted exceptions, the central dogma motivates the collection and integrative analysis of intermediate molecular phenotypes from the transcriptome, proteome and metabolome (also depicted in figure 1). Last is the concept of hidden structures which is important due to the many unobservable hidden states of biomolecules. A very large class of modeling situations can be represented as having a part that can be observed and part that cannot be observed. However, inference about the unobservable part can be done since it interacts with the observable part.

C1 A Mapping from Genotype to Phenotype

C2 Networks

C3 Genealogical Relationships

C4 Biological Knowledge

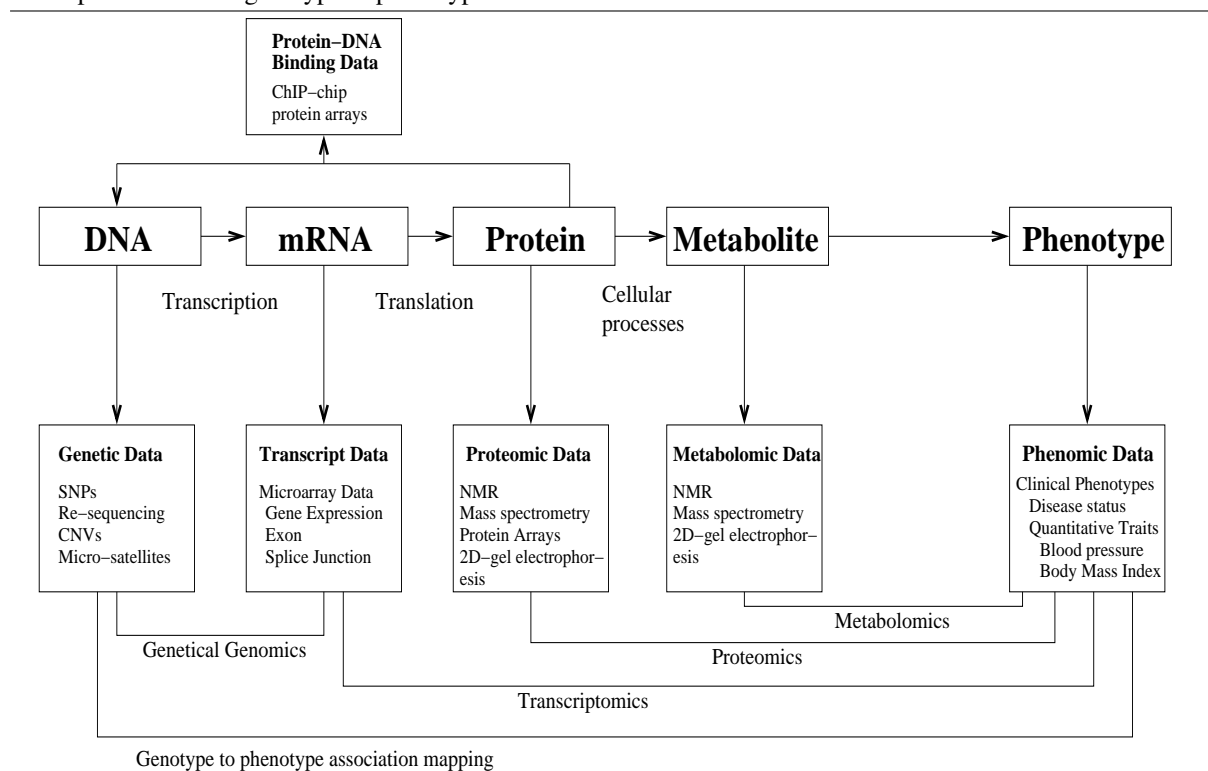
C5 Hidden Structures

Data and concepts are combined to construct models. These can be used to interpret and analyse data, and different combinations of data, knowledge, and concepts lead to different analytical techniques. Most reasoning and communication between researchers is done via discussion of models, hence the importance of models can hardly be exaggerated. Choosing and constructing appropriate models that describe data in the presence of noise requires the use of statistics and is the focus of section 4. We evaluate a range of analytical techniques which range from the well established analyses of single sources of biological data through to more recent and complex integrative strategies.

The goal of studies which analyse global biological variation data is usually to refine knowledge and understanding of biological processes, usually with respect to a particular phenotype or certain conditions. In particular, the most recent integrative analyses of multiple data types aim to suggest putative causal mechanisms underlying a phenotype. Complete molecular descriptions of these mechanisms can be ascertained using systems biology, but this approach can only be used on a small scale relative to the global studies of variation which we describe a framework for in this paper.

As an increasing number of studies of global variation provide evidence which support putative causal mechanisms giving rise to complex phenotypes, the field over the coming years will become increasingly focused on finer scale functional studies which we introduce and discuss in section 5. At present there are few functional studies which have completely described the biological mechanism causing a complex phenotype. Most at best can suggest possible explanations, which is not surprising given that a complete functional dissection will require detailed knowledge and accurate dynamic modelling of many biomolecules.

Figure 1 The main biological mechanisms involved on the path from genotype to phenotype, the biological data types harvested to capture these processes together with various disciplines which have evolved to analyse data. Note that epigenetic and environmental exposures are missing from this figure and also play a vital role in many of the processes from genotype to phenotype.



2 Data Types

There are six main sources of data which can be measured on a global scale and in this section we distinguish between the true source of data (e.g. transcripts, proteins, DNA, metabolites) and the typically measured quantities representing them. We discuss the completeness, dimension, and reliability of such data, as well as other notable features such as correlation patterns.

2.1 Genomic Data

Full genomic DNA data are available at a species level with the total number of organisms fully sequenced ever increasing [99]. The sequenced human genome was largely completed in 2001 [132, 62, 63], a consensus of approximately three billion nucleotide positions achieved using only a few individuals. Obtaining the entire genomes of single individuals is now a possibility [82, 138], and a large-scale project is forthcoming to identify the six billion basepair diploid genomes of roughly 1000 individuals (<http://www.1000genomes.org/>). Though DNA can vary within an individual from cell to cell, currently a consensus is taken across cells (often from blood samples) to achieve a single representation of an individual's genome.

However, the cost of exhaustively collecting all of an individual's genetic information is currently too high to be done in the vast majority of studies. Instead, genetic information is usually collected at a subset of genetic *markers* that attempt to capture as much of the complete genome information as possible. Approximately 99% of the DNA between any two individuals is the same. Markers are genetic regions that show variability in a population. There are three particularly prominent types of genetic marker:

1. Single-Nucleotide-Polymorphisms (SNPs) – nucleotide positions where a sampled chromosome can take different values, or allelic types. SNPs are the most common type of genetic marker in the human genome, comprising ≈ 0.1 -1.0% of the genome in most human populations. The vast majority of these have two allelic types.
2. Micro-satellites – regions of contiguous repeated short nucleotide sequences where a sampled chromosome can have a different number of repeats, comprising <0.1 of the genome. Each micro-satellite location thus often has more than two possible allelic types.
3. Copy-Number-Variants (CNVs) – 1Kb to several megabase segments of the genome that are present in different numbers in different chromosomes sampled from a population. Recent work suggests that $\approx 12\%$ of the genome is subject to copy-number-variation [106].

Most large-scale studies currently collect genetic information at ≈ 300 -1000K genetic markers, typically SNPs. The rate of mis-classifying the allelic type at a genetic position depends on the type of marker analysed and the method used to classify it; for SNPs this rate is typically $<1\%$ (e.g. the estimated error rate was ≈ 0.1 -0.2% in the WTCCC study).

Though markers consist of only a subset of the genome, there is often a considerable amount of correlation between them. This phenomenon is known as Linkage Disequilibrium and occurs because genetic material at distinct loci are not transmitted independently (see section 4.2.1). SNPs typed on genotyping arrays are carefully selected such that they can potentially capture a much larger proportion of the total genetic variation.

In typical genetic data collections, one collects only allelic information at each marker and not information on which alleles were inherited together along a chromosome from a common source. For example, in diploid organisms such as humans, an individual has two chromosomes representing a given region, one inherited from each of their parents, so that each marker of an individual is a set of two observations. However, considering two consecutive markers, in typical DNA sample collections we do not observe which of the two possible combinations of alleles across markers were inherited together from each source. This is known as the *haplotype information*, which can potentially be estimated using existing software exploiting linkage disequilibrium patterns. Similarly, imputation algorithms [85] can be used to infer genetic variants that are not directly typed on an array or even to predict misclassified positions.

2.2 Epigenomic Data

Epigenomics is the study of epigenetics at a global scale. There are a variety of definitions of epigenetics [9], but we define it as the study of global features/processes which have influences on cellular regulation which are not encoded by DNA sequence variation. The main areas of study are chemical modifications to DNA and structural changes in DNA packaging, in particular DNA methylation and histone modification. Methylation is more stable, so changes in methylation patterns are typically longer lasting than histone modifications which might only last for the duration of the cell life.

Methylated DNA (mainly methylated cytosines) and histone modifications can be detected using chromatin immunoprecipitation techniques. There are specific antibodies against 5-methyl cytosine which can be used to immunoprecipitate methylated DNA fragments and similarly there are specific antibodies against various modified histones which can be used to immunoprecipitate the modified histone proteins. These fragments can then be washed, amplified and hybridised to microarrays. Bisulphite treated DNA can also be used to detect methylated C's since the treatment converts unmethylated C's to T's. Bisulphite DNA can be hybridised to an array to distinguish methylated positions. To reduce the dimension and the number of arrays required (for both ChIp-on-chip and Bisulphite treatment methods), the arrays used usually contain probes from gene promoter regions.

Epigenetic processes are widespread throughout the genome. Epigenetic features are primarily inherited but they are subject to modification over time; either due to environmental sensitivity or stochasticity associated with inaccurate copying mechanisms. The effects of an epigenetic change/mutation can be passed onto daughter cells (e.g. cancer tumour growth) or be temporary and last only a single cell cycle. For example, the copying mechanisms associated with DNA methylation for example are only 96% accurate such that one error is expected every 25 methylated sites. Since the epigenome of an individual is dynamic over time, it is suggested that it could be responsible for incomplete penetrance of genetic disease. For example studies show that identical twins can exhibit vast differences at an epigenomic level and hence could be underlying discordant phenotype.

The study of global epigenomics is not yet widely employed, although advances in technology are making it increasingly viable. Instead the study of epigenetic processes is usually reserved for more focused studies. Consequently we discuss the utility and applications of these sources of data in section 5.

2.3 Transcriptomic Data

Transcription is the process by which RNA is synthesised from DNA in the nucleus of the cell. It can be considered in two stages (see Figure 2). First proteins (transcription factors) bind to DNA to induce transcription of primary messenger RNA transcripts (i.e. *pre-mRNA*) from DNA. The second stage, RNA processing, involves the splicing of *pre-mRNA* transcripts into mature mRNA (i.e. *mRNA*). Splicing is the process by which exon transcripts are separated from intron and intergenic transcripts and joined back together. Since exons can be included or excluded in different ways, multiple transcript isoforms can be made following the initial transcription of a single gene, a phenomena is known as *alternative splicing*. Thus while the entire genome may theoretically be transcribed into *pre-mRNA*, only the exons of genes may be transcribed into the mRNA that leaves the cell to synthesize proteins. Other mechanisms also influence translation of mRNA into protein so there might be little correlation between observed quantities of mRNA and the analogous protein. Despite this, mRNA abundances serve as an indirect measure of gene expression or activity.

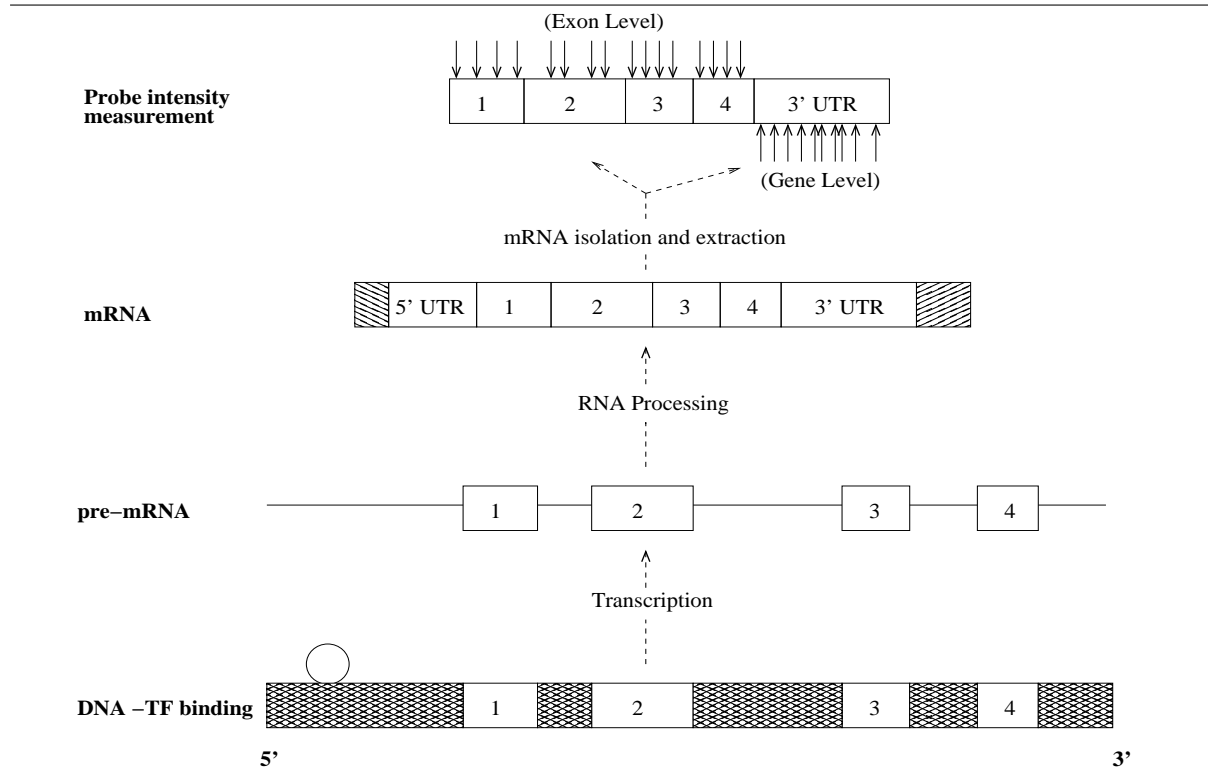
Mature RNA levels vary between different cells from the same source. The expression of different genes allows cells to differentiate and perform different functions. Consequently, unlike genetic material which is nearly identical throughout all cells in the same individual, genes expression patterns vary according to cell type (and there is stochasticity even within cell type) and are dynamic in time. Single cell techniques target mRNA level within individuals cells, but the majority of studies involving large numbers of individuals do not use such a fine resolution. Instead, microarray-based techniques are commonly used which interrogate thousands of mRNA levels simultaneously from a sample of a large number of cells ($\approx 10^3$), and final observations are therefore averages rather than cell specific observations.

There are approximately 24,000 known human genes [25], whose mRNA transcript levels can now be routinely interrogated using a single microarray which is often referred to as a gene expression array. The mRNA transcripts present in a sample are quantified via the use of many short oligo-nucleotide probes (see Figure 2). Usually

between 4-11 probes comprise a probe set and each probe set targets a region of a specific gene. There may be several probe sets which target the same gene according to its length. Microarrays designed to target multiple different exons throughout a gene can have upwards of 1 million probe sets each containing 4 probes. Intensities observed at probe sets within a single exon provide a measure of expression of all transcripts which contain that exon (which may be many due to alternative splicing). However, standard expression microarrays do not provide exon level data, with probe sets located only towards the 3' untranslated region (UTR) of the gene. There are fewer probe sets in standard arrays, but they each typically contain more probes (up to 11). High correlations of transcript abundance measures are expected between probe sets which map to the same gene and indeed, other patterns of correlation are also expected according to the complex regulatory systems operating to control gene expression.

Most analyses summarise each probe set by a single value (typically a robust average across probes). These values do not map 1-1 directly to mRNA levels, due to technical and biological sources of noise. Noise can be introduced by variability across different lab technicians, array stickiness or fluorescent dye effects, and there are varying degrees of non-specific and specific hybridisation i.e. transcripts other than the target can also bind to the probe, and some targets bind more readily than others. Consequently, probe intensities are not interpreted directly in terms of the number of transcripts present in a sample. Instead subsequent analyses consider intensities relative to one another. Suitable pre-processing of the raw data enables within array and between array comparisons to be made.

Figure 2 An illustration demonstrating how mRNA transcript abundance is quantified for a protein coding gene. In the bottom layer, transcription of the gene is initiated by the binding of a transcription factor (the circle) to DNA and produces pre-mRNA. Protein coding exons are numbered and introns are filled. The resulting single stranded pre-mRNA is illustrated in the second level with non-coding regions represented by a single line. The third layer illustrates RNA processing (i.e. splicing, capping at the 5' end of gene and the addition of the poly-A tail to the 3' end of the gene) and results in the formation of mRNA which is transported outside of the nucleus of the cell. Different isoforms can be produced according to alternative splicing although only one isoform is illustrated in this figure. Finally, mRNA is isolated, extracted and hybridised to a microarray containing probes for the transcript. This is demonstrated in the top layer of the figure by the arrows. Each arrow represents a single probe. They are usually short oligonucleotides approximately 25mers. At a gene level, probes are located in the 3' untranslated region (UTR) whereas at an exon level, probes are distributed throughout exons of the gene. At the exon level, probes within a single exon form a probeset, in this example there is a single probeset per exon although this is not always the case. At a gene level, in this example there is only one probeset but similarly, there may be multiple probe sets depending on the length of the gene.



In addition to mRNA, there are several types of non-coding RNA, which are not translated to protein. Ribosomal and transfer RNAs (rRNA, tNA respectively) are crucial for the translation of mRNA into protein in every cell, while micro-RNAs (miRNAs) have a regulatory role. miRNAs are shorter than regular protein coding RNAs (\approx only 100 bases processed to \approx 21-23 nucleotides) and they are approximately complementary to mRNA. This enables them to bind to mRNA and interfere with post-transcriptional processes and protein synthesis hence regulating protein gene expression. Since they need not be perfectly complementary, a single miRNA can regulate multiple genes. There are microarrays designed to target miRNAs specifically. There are approximately 800 known or putative human miRNAs which can be screened for simultaneously using similar microarray technology.

2.4 Proteomic Data

Proteins are synthesised from mRNA transcripts via translation and interact with each other and other biomolecules to perform a wide range of functions within a cell. The activity level or expression of a protein coding gene is more directly measured by quantifying the amount of synthesised protein. The total size of the human proteome is

estimated to be at least ten times greater than the total number of protein coding genes ($\approx 24,000$), with the space of potentially physiologically relevant protein-protein interactions recently estimated to be $\approx 650,000$ [124].

Like mRNA transcripts, protein abundances cannot be measured directly and, in the same way that a single gene transcript expression is targeted by multiple probe sets, protein presence and abundance is targeted via the identification and detection of small peptides which make up the protein. Single cell global protein profiling is not yet viable. Consequently, as in transcript profiling, it is subject to averaging of abundances over a sample of cells, thereby neglecting local dynamic and stochastic information. Here we describe the main techniques for detecting proteins on a global scale. Often, techniques to establish structural properties of proteins (such as phosphorylation and methylation) are reserved for more focused studies and are not used on a global scale.

There are several types of technology which can target different proteins and their structural properties. We concentrate on those which are used on a global scale (i.e. protein identification and abundance) and we describe the main data types which are also reviewed in more detail by Chaerkady and Pandey [18]:

- **Mass Spectrometry** - The proteins in a sample are subject to chemical fragmentation resulting in charged peptides. The charge and mass of these peptides vary according to their composition hence enabling protein identification. The data is usually presented as a graph with the mass-charge ratio of a particle along the x-axis and abundance on the y-axis. The identification of proteins requires deconvolution of this data. There may be multiple 'peaks' in the graph corresponding to the same protein since they get fragmented into shorter peptides.
- **2D gel Electrophoresis**- This technique separates proteins in two orthogonal dimensions on a gel. They are separated according to specific properties of the protein such as isoelectric point or protein mass. Up to 10,000 protein spots can be separated in a single gel [71] and the resulting data are images of the gels with statistical methods used to identify unique spots. Noise (i.e. weak spots) and incompletely separated spots require careful processing.
- **Protein Arrays** - arrays are spotted with protein specific antibodies, aptamers or affibodies. Targeted protein abundances are quantified upon hybridisation of the sample to the array. [56]

Since proteins physically interact with each other and other molecules (such as RNA binding) in cellular processes, correlated abundances are to be expected. Protein interactions are usually inferred via observations of direct physical interaction but other correlations can be observed from abundance data. Mass spectrometry data has further correlation structure to the multiple 'peaks' corresponding to peptides derived from the same protein.

2.5 Metabolomic Data

Metabolites are the products of cellular processes involving proteins and/or other metabolites. Examples include carbohydrates, fatty acids and amino acids. They can be endogenous to the metabolism or exogenous compounds such as drugs, and are therefore sensitive to genetics, epigenetic processes and environmental exposures. Consequently, the identity, abundance and structure of metabolites present in a sample can be informative about cellular regulation perturbed by or driving phenotype and disease.

The size of the metabolome and what precisely constitutes a metabolite is still a matter of debate. Currently there are ≈ 6500 categorized human metabolites (<http://www.hmdb.ca/>), though the total number of human metabolites may be on the order of tens of thousands. The range and number of metabolites detectable depends on that platform used for quantification, such that complete metabolic profiling requires interrogation using a variety of techniques. Nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), chromatography and vibrational spectroscopies are amongst the most widely employed and are reviewed by Dunn and Ellis [34]. The output from MS studies is similar to that of protein mass spectrometry data; a mass spectrum resulting from the ionising and separation of a sample. The data can be represented using a graph with mass-charge ratios along the x-axis and intensities along the y-axis. Statistical analysis is required to extract and identify the 'peaks' with specific known metabolites or classify them as novel molecules. The number of peaks identified varies according to the study but they are often of the order of hundreds. The output from NMR spectroscopy is also data which forms peaks which correspond to metabolites. This technique can also be informative of the metabolite structure in addition to intensities according to the metabolite abundance. Dunn and Ellis [34] provide further detail and

discuss the merits of different platforms. The platform selection should reflect the objectives and the processes of primary interest since they vary in sensitivity, technical noise levels and molecular targets.

Metabolic profiling can be done with a variety of different samples including intact tissues and biofluids. Functional interpretation of biofluid metabolic profiling is complicated since metabolic reactions and the processes by which they are produced can be influenced by multiple organs, tissue and cell types. However, as in transcriptomics and proteomics, metabolite profiles are currently assessed using a sample of cells rather than on a cell-by-cell basis. Furthermore, the metabolic behaviour is more dynamic than either that of mRNAs or proteins and can adapt quickly in response to environmental changes.

Metabolites can be the product of a process involving proteins or they can result from chemical reactions involving other metabolites and specific catalytic enzymes. This induces correlation between the abundances and presence of different metabolites. For example the presence of metabolite C might depend on the presence of metabolites A and B and enzyme E.

Correlations can be observed from static observations of metabolite abundances can be considered representative of the overall state of a system but they are sub-optimal for determining sets and rates of biochemical reactions which are clearly dynamic. Systematic perturbation experiments coupled with time series metabolic observations provide more information although sampling at a suitably fine time scale is not always possible. In-vitro biochemical experiments can also generate informative metabolic data, although there may not be correspondance between in-vitro and in-vivo measurements.

2.6 Phenomic Data

Phenomics is the study and characterisation of phenotype and by definition the term *phenotype* can be used to describe any observed manifestation of genotype. Ideally, global phenotyping of any individual should include many measurements to cover a wide range of characteristics including those which are morphological, biochemical, behavioural or psychological [42]. Furthermore, there should be standardised procedures in place such that measurements are comparable across countries. In practise global scale phenotyping in human individuals remains an ideal (with the human phenome project proposed in 2003 [42]) but there have been considerable advances towards achieving this goal for mice [12]. In particular, there are now centres throughout different countries where mice can be sent to be ‘phenotyped’ according to standardised tests.

In most human studies there are relatively few phenotypes catalogued which target morphological, behavioural and psychological traits. The majority are phenotypes which are traditionally of medical origin and include disease diagnosis, presence or absence of a symptom, body mass index, blood pressure and response to skin prick tests. Precision and dimension of phenotype characterisation has not improved in line with genetic, transcriptomic, proteomic and metabolic profiling. This is mainly due to the subjective nature in which they are ascertained. For example different clinicians might ask different questions regarding symptoms of the disease. The International Statistical Classification of Diseases and Related Health Problems (ICD) defined by the World Health Organisation (WHO) provides a way of systematically classifying and coding every health condition according to six categories. However, since different molecular diseases or phenotypes can manifest in similar or the same symptoms, accurate and complete phenotyping remains a challenge. Measurements/observations can be incomplete, subjective and imprecise and yet they form the basis of many studies. Biochemical phenotyping or molecular phenotyping for humans is more comprehensive and includes global observations of transcripts, proteins and metabolites which as discussed previously can be measured (albeit indirectly) according to rigorous protocols worldwide.

Strictly speaking, the human phenome encompasses all observations of (E, T, P, M) and arguable even G. However most researchers use the term phenotype to describe disease status, which is a very coarse observation. The task of well defining appropriate phenotype(s) in a study of the mapping from genotype to disease is fundamental to its findings and success with respect to sensitivity (detecting true positives) and specificity (rejecting true negatives). For example, Posch *et al.* [100] call for supplementary phenotypes to distinguish between congenital heart disease conditions of different etiologies, and Bilder [8] suggests that the lack of significant genetic association findings for psychiatric diseases such as bi-polar disorder and schizophrenia may be due to the sub-optimal nature of categorical psychiatric diagnoses as phenotypes.

Correlation between molecular traits are expected as discussed previously, however, it is rare that multiple phenotypes/disease states are ascertained from the same study individuals, hence disease correlations are not well

understood. There are some well known instances of co-morbidity of disease including asthma with eczema, and obesity with type 2 diabetes. In other situations, the presence of one disease or phenotype can be negatively correlated with another. For example, sickle cell anaemia is protective for malaria and prostate cancer is protective for type two diabetes (and vice-versa). Correlation between clinical phenotypes are usually caused by a degree of overlap of disease risk factors of which some might be genetic.

2.7 Other Data Sources

The data types G, T, P, M, E and F are not comprehensive, in this section we briefly describe some other sources. The majority are image based and reserved for focused studies rather than phenotype studies of large numbers of individuals. They can be categorised according to their target which are either located within the cell, beyond the cell (and within individual) or beyond the individual (external environmental quantities).

Within the cell quantities include G, T, P and M, but data types as described thus far largely quantify presence/absence or abundance data. Structural properties and dynamic behaviour can also be interrogated albeit on a smaller scale. Crystallography, electron cryomicroscopy (cryo-EM) and microscopy can provide structural information on molecules, and real time dynamic behaviour of labelled positions on a single molecule can now be determined using ‘Single Molecule Measurements’ [75, 73]. They are largely optic based and can give detailed information in the dynamics of movements of for instance parts of RNA, proteins or molecules in a membrane. They provide data at the most detailed level and contrary to other techniques do so without any need for averaging.

Biological phenomena which extend beyond the cell are of paramount importance for the study of complex phenotype and they are essential to describe how perturbations within a cell or population of cells manifest in observed phenotypes. Information can be obtained via observations of cellular quantities in different tissues for example but few studies are able to employ this approach due to both expense and practicality of sampling relevant cells in large numbers of individuals. The majority of observations ‘beyond the cell’ are imaging techniques and include confocal microscopy, magnetic resonance imaging (MRI), tissue image cytometry and optimal projection tomography which can provide four dimensional high resolution data at an organ level. Observations from these sources contribute to understanding the signalling and processes which transfer perturbations in the cell to perturbations of phenotype.

Environmental exposures are perhaps one of the most difficult factors contributing to complex disease to study. Even if they were known, monitoring the human environment and collecting data is difficult. Obvious factors include smoking, diet, alcohol intake and stress but even they cannot be measured to a high degree of accuracy. For this reason, many gene-environment studies are performed with model organisms which can be subjected to homogeneous conditions. In humans, to some extent, the effects of environmental influences might be reflected within the individual but distinguishing these effects and attributing them to external factors is difficult.

ARRAY/PLATFORM TYPE	TARGET	DESCRIPTION	DIMENSIONALITY
Genotyping	DNA; SNPs, CNVs	Captures genetic variation at SNPs genome-wide. CNVs can also be inferred.	Up to $\approx 1/30$ of all SNPs ($\approx 75\%$ of genetic variation) can be typed on a single array
Re-sequencing	DNA; Continuous lengths of DNA.	Captures all forms of genetic variation including SNPs, CNVs, Chromosomal rearrangements and rare variants.	Up to 28 probes for two stranded re-sequencing per nucleotide.
Comparative Genomic Hybridisation (CGH)	DNA; CNVs	Captures copy number variation of regions of the genome by comparison with a reference sample. Micro-deletions and amplifications can be detected.	Resolution is determined by spacing and probe length; for genome-wide arrays, probes are equally spaced \approx one per Mb.
RNA transcript	RNA; transcripts	Captures global 'gene expression' by measuring relative mRNA transcript abundances.	Up to $\approx 54,000$ probesets for mRNA transcripts per array.
Exon/ splice-junction	mRNA; transcripts	Probe sets contain fewer probes but there are more relative to regular 'gene expression' arrays. They are distributed throughout exons of a gene and across splice junctions.	≈ 4 probes in each probeset, one or more probesets per exon, yielding a total of 1.4 million probesets
miRNA arrays	RNA: miRNA transcripts	Captures expression of short non-coding RNAs.	Probe sets for ≈ 500 of the known miRNAs (in human) and an additional set of putative miRNAs on a single array. (check with Q)
Protein Arrays	Proteins/peptides	There are several types of protein array including antibody arrays, peptide arrays and protein-DNA arrays they all simultaneously quantify protein expression of specific proteins.	The number of target proteins varies according to the array type.
ChIP-on-chip	Protein-DNA interactions	Chromatin immunoprecipitation (ChIP) is used to isolate a specific protein and its bound DNA. The DNA is mapped to the genome via hybridisation to an array (on-chip). Applications include the detection of specific DNA-protein binding sites, methylated sites and structurally modified proteins.	Usually tiling arrays are used an the number of probes vary according to the resolution. Multiple high density arrays required per human chromosome.
ChIPSeq	Protein-DNA interactions	Chromatin immunoprecipitation for protein isolation, followed by sequencing of the DNA to map protein binding DNA to the genome.	Short lengths of DNA
Methylation arrays	DNA; Methylated Cytosines	An alternative to ChIP-on-chip, since methylated cytosine's remain unchanged by bisulfite treatment, sulphite DNA can be used to detect methylated positions.	Short 25mer probes at each (candidate) methylated site. Analysis is usually restricted to promoters of specific genes
2D-gel Electrophoresis	Protein	Proteins are identified and quantified via separation of the molecules into orthogonal dimensions; typically the isoelectric point and protein mass.	
Mass Spectrometry	Protein	Various types, all based on the mass to charge ratios of ionised protein molecules which are used to infer the true mass of the molecule. Proteins/peptides can be uniquely identified if their exact mass is known.	The dimensionality depends on the type of technology and any previous filtering of the proteins. Often multiple peptides per protein.
NMR			

Table 1: Summary of available bio-technologies and the types of high-throughput data they generate.

3 Concepts

Interpretation and analysis of observed data requires the formulation of appropriate models and these are founded on some general concepts relating to the structure, dynamics and evolution of an organism. There are five main concepts which we discuss in this section: a mapping from genotype to phenotype, networks, genealogical relationships, biological knowledge and hidden structures. Genotype to phenotype functions are very general, they rarely attempt to describe functionality but instead focus on predicting the modification to disease risk or phenotype in the presence of different genetic variants. Networks are again models, but they attempt to provide a more functional explanation by involving quantities that can be interpreted at the molecular level. Contrary to genotype to phenotype functions and networks which both provide approximations to true mechanisms, true genealogies relating cells, individuals and species exist and could be exploited effectively if they could be observed. In practise, this is rarely possible, so models of evolution are important to characterise the uncertainty over possible genealogies consistent with the data. All studies are founded on a certain level of biological knowledge and as research studies continue to exploit high throughput experimental techniques, representation and exchange of knowledge is becoming increasingly important. Hidden structures have also played an important role in data analysis and are discussed in the last part of this section.

3.1 A Mapping from Genome G to Phenome F

Figure 1 depicts a series of processes connecting the genome G to the phenome F , but, much work in the past several decades has concentrated on associating genetic variation data directly to a phenotype (largely because genetic data was the first to become readily available in large quantities). Most studies simplify the mapping by considering mapping a single phenotype $y \in F$. An increasing number of studies are investigating multiple phenotypes, such as molecular or anthropometric traits but these are typically mapped to the genome (or separate loci in the genome) separately. Hence we primarily discuss mappings from the genome to a single phenotype in this section (equation 1). The functions $h(\cdot)$ and $f(\cdot)$ are mapping functions and $\mathbb{E}(\cdot)$ denotes expectation in the statistical sense. Most mapping functions model the expectation of a phenotype (i.e. $\mathbb{E}(Y)$) which accounts indirectly for noise/ unknown sources of variation. These mapping functions include the well known class of statistical models; Generalised Linear Models (GLMs).

$$h(\mathbb{E}(Y)) = f(G) \quad (1)$$

Mapping the genome to a single phenotype is done by breaking down the genome into regions according to a set of genetic markers or simply by mapping a subset of genetic loci which show variation in a population. The genetic variation at a single locus is denoted by g and is usually quantified by a single value which aggregates the variation on both parental chromosomes (for diploid organisms). In the case of biallelic SNPs which take one of two types on each chromosome, variation is quantified numerically according to the number of alleles of (labelled arbitrarily) one of the two types, hence genotype values can be represented by 0,1 or 2. Hence g is commonly a value in 0,1,2 for each locus. Naive mapping functions consider mappings from genotypes at single genetic markers to single phenotypes (equation 2).

$$h(\mathbb{E}(Y)) = f(g) \quad (2)$$

Although this class of functions can suffice for phenotypes influenced by a single gene they are often inadequate to describe genetic influences on complex phenotypes which might be better described by functions involving multiple markers (equation 3 where there are m loci affecting y).

$$h(\mathbb{E}(Y)) = f(g_1, \dots, g_m) \quad (3)$$

Mapping genetic markers or genomic regions to a phenotype involves characterising how variation at markers (single or multiple) influences phenotype or risk. This involves defining a penetrance function which makes statements about phenotype conditional on genotype at a particular locus (or set of loci). A simple example of a penetrance function (and indeed that of a map from genotype to phenotype) might say that if specific genotype is

present at a particular locus then an individual has a particular (disease) phenotype with probability 1 (equation 4). Note that equation 4 is an example of equation 2 where $h(y) = \mathbb{P}(y = disease)$ (or equivalently $\mathbb{E}(Y)$ if disease is encoded 1 and control is coded zero) and $f(g_1)$ is an indicator variable which takes the value 1 if $g_1 = g$ and zero otherwise. Genetic effects are highly or completely penetrant for mendelian phenotypes such as sickle cell disease.

$$\mathbb{P}(y = disease) = I_{\{g_1=g\}} \quad (4)$$

Incomplete penetrance is the term used to describe instances when phenotype is modified with probability less than 1 in the presence of a genetic variant. Incomplete and low penetrance is common for complex phenotypes and is reflective of other influences on phenotype such as other genetic, epigenetic or environmental exposures. That is, the presence of specific genotype at a locus might modify phenotype only if it is activated in some way by another genetic variant, methylation pattern or smoking for example. Hence the probability of the phenotype being modified given the genetic variant is present depends on the prevalence of the other factors which interact with the locus. If observations of these factors are available then they can be included into the mapping function (equation 5). In this equation e denotes epigenetic factors and x denotes external environmental exposures. These may also be multi-dimensional.

$$h(\mathbb{E}(Y)) = f(g_1, g_2, \dots, g_m, e, x) \quad (5)$$

The class of mapping functions in equation 5 is vast. The simplest approach is to consider the class of functions which describe main additive effects (equation 6).

$$f(g_1, g_2, \dots, g_m, e, x) = \sum_{i=1}^m f_i(g_i) + f_{m+1}(e) + f_{m+2}(x) \quad (6)$$

However this class of functions does not model genetic, environmental or gene-environment interactions. Clearly, with a large number of possible genetic, epigenetic and external environmental factors the function space encompassing all interaction models it is not possible to consider exhaustively. Consequently the interaction mapping functions are usually restricted to include pairwise or low order terms. Examples are equations 7 and 8 where there are separate terms for main effects and interactions.

$$f(g_1, g_2) = f_1(g_1) + f_2(g_2) + f_I(g_1, g_2) \quad (7)$$

$$f(g, e, x) = f_1(g) + f_2(e) + f_3(x) + f_4(g, e) + f_5(g, x) + f_6(x, e) \quad (8)$$

A statement about the mode of inheritance of a phenotype can be implicitly defined by the mapping function, that is, how phenotype is transmitted from one generation to another (assuming no de novo mutations). This is difficult to define for complex phenotypes affected by multiple genetic loci and other factors, but for single gene/locus phenotypes, dominant, recessive and additive modes of inheritance can be well defined and modelled. Dominant genetic effects on phenotype are those which require only one risk allele to modify phenotype (i.e. inherited from either or both parents), recessive effects require the risk allele to be present on both chromosomes (i.e. inherited from both parents) and additive effects see the presence of each risk allele modifying phenotype or some function of the phenotype additively. Dominant and recessive mapping functions can be specified with the previous equations by re-coding genotypes to 0 and 1 according to whether or not a single or two risk alleles are present. For recessive models, genotypes 0 and 1 are both recoded to zero and genotype 2 is recoded to 1, where as for dominant models, genotypes 1 and 2 are recoded as 1 and genotype 0 remains the same.

Genotype to phenotype mapping functions are widely used throughout studies of phenotype and in table 2 we describe the main classes of functions. These examples belong to the class of Generalised Linear Models (GLMs) where in each case $h(\cdot)$ is a function of $\mathbb{E}(Y)$ (where Y the phenotype is considered to be a random variable and y is an observed value of the random variable). For linear mapping functions, $h(\mathbb{E}(Y)) = \mathbb{E}(Y)$ and for logistic mapping functions, $h(\mathbb{E}(Y)) = \log \left[\frac{\mathbb{E}(Y)}{1-\mathbb{E}(Y)} \right]$. Note that some analyses do not use a function $f(\cdot)$ at all and

merely seek to establish whether an informative map exists. A prominent example is in genome-wide association studies, where chi-squared contingency tables that test for independence between variants at a genetic locus g and case/control status (i.e. $y \in$ (“diseased”, “non-diseased”)) are often used to measure associations between genotype and phenotype [6]. In a similar way genetic interactions influencing a disease phenotype can be tested without the use of mapping functions or models; under the assumption that there is no interaction between a pair of loci (or higher order interactions) the probabilities of observing a disease phenotype given the genotypes at both loci decomposes as a product of the appropriate probabilities. Hence a chi-squared contingency table can be constructed in a similar way.

FUNCTION CLASS	FUNCTION	DESCRIPTION
Linear functions	$\mathbb{E}(Y) = \alpha + \beta g$	Single marker model
	$\mathbb{E}(Y) = \alpha + \sum_{i=1}^m \beta_i g_i$	Multi-marker model
	$\mathbb{E}(Y) = \alpha + \sum_{i=1}^m \beta_i g_i + \sum_{i \neq j} \beta_{ij} g_i g_j$	Pairwise interaction model
Logistic functions	$\log \left[\frac{\mathbb{P}(Y=1)}{(1-\mathbb{P}(Y=1))} \right] = \alpha + \beta g$	Single marker model
	$\log \left[\frac{\mathbb{P}(Y=1)}{(1-\mathbb{P}(Y=1))} \right] = \alpha + \sum_{i=1}^m \beta_i g_i$	Multi-marker model
	$\log \left[\frac{\mathbb{P}(Y=1)}{(1-\mathbb{P}(Y=1))} \right] = \alpha + \sum_{i=1}^m \beta_i g_i + \sum_{i \neq j} \beta_{ij} g_i g_j$	Pairwise interaction model

Table 2: Examples of commonly used mapping functions for continuous and binary phenotypes. Linear functions map quantitative traits furthermore if errors are normally distributed these are ordinary linear regression models. Logistic functions map categorical phenotypes including disease case-control status. Both of these are examples of GLMs.

Most functions including those described in table 2 are based on minimal biological knowledge, though there are exceptions (see Figure 3.1). Much of systems biology can be seen as attempts to create genotype to phenotype functions based on functional knowledge. Ideally, one should take a standard genome with a standard model of the organism and predict the result of a change in the genome. This approach whilst ideal, in practise is severely restricted by the lack of biological understanding about processes. For example it was only relatively recently that RNA interference was discovered as a mechanism contributing to gene regulation. Integrative approaches may well help in this regard, providing further knowledge to predict models that might be useful/appropriate for representing complex systems. However with current data precision and resolution such approaches are also limited. It could be that increasingly ambitious modelling could reveal hidden components and lead to fundamental discoveries. If this will be the case remains to be seen, particularly since simple models (even if they are wrong) are often preferable for biological interpretation.

3.2 Networks

The concept of network is extremely general; it is a set of objects with a set of relationships. Relationships can also be a set of objects and both these objects and relationships can be labelled. Thus the ubiquity of networks is not surprising, but can they describe everything?

In most cases, networks can be described using graphs which have a precise mathematical definition in terms of a set of (labelled nodes) and a set of edges (specified by pairs of nodes). Edges can also be directed or undirected.

3.2.1 Biological Networks

A biological network is a network used to describe some kind of biological system, for example, a cell the size of *E.coli* with $\approx 10^9$ - 10^{10} molecules might be represented by a network with say 10^3 - 10^4 nodes and edges labelled using molecular data. Modelling human biological systems involves additional levels of complexity; each cell has approximately 10^{13} atoms and there are $\approx 10^{13}$ cells, furthermore a full dynamic description at an atomic level requires $\approx 10^{15}$ time steps per second, hence a complete description of the human system is of order $\approx 10^{41}$ atomic positions per second. Human systems are decomposed according to different tissues, cells and processes yet even these subsystems are (at most) represented by networks involving $10^3 - 10^5$ nodes and edges. The reduction of at least 36 orders of magnitude is founded on substantial molecular and temporal approximations:

1. Molecular Approximations:

- Biomolecules represented by their observed abundance e.g. a gene represented by its observed mRNA expression level.
- Nodes (labelled with genes for example) considered 'on' or 'off'.
- Physical interactions between molecules considered to be 'present' or 'absent'.
- Many molecules excluded, either because they are unobserved or not considered important to the system being modelled.

2. Temporal Approximations:

- Single snap shot observations of data to construct networks representative of a system at a single point in time (usually assumed to be in a steady state).
- Dynamical systems approximated by a few characteristics such as rate parameters in a system of ordinary or stochastic differential equations.
- Dynamical systems approximated according to observations at a discrete set of time points appropriately chosen according to the time scale of the system of study.

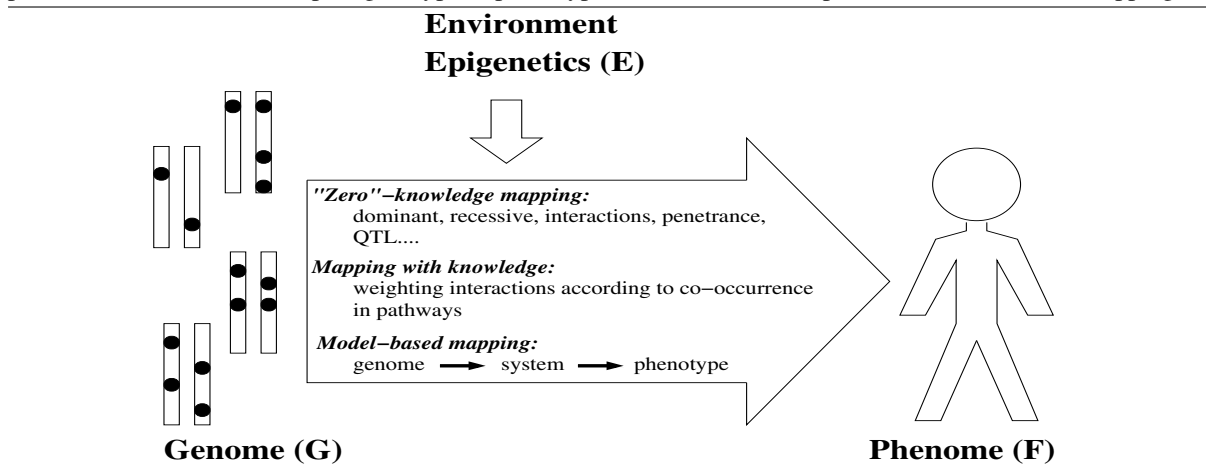
It is not surprising therefore that resulting networks can rarely be representative of a complete physical system. Network approximations are further limited by the type and precision of data used to construct them. Often (particularly with measurements of proteins, transcripts and metabolites), the target is measured indirectly and subject to multiple sources of noise as described in section 2. How valid and network approximations and data measurements are will only be revealed as increasingly ambitious attempts are made to simulate and accurately characterise simple dynamical systems.

There are four well established types of biological network (N1-N4) which aim to approximate the processes determining function and phenotype at a cellular level.

N1 *Protein Interaction Networks* - Nodes are labelled with proteins and edges are representative of a physical interaction.

N2 *Signal Transduction Networks* - Nodes are labelled with signals, e.g. hormones, proteins and edges represent biochemical processes by which the signals are transferred and converted to other signalling molecules. Since the processes occur in sequence the resulting networks are often referred to as cascades.

Figure 3 Mapping the Genome to the Phenome. This mapping can incorporate standard genetic concepts but still not assume anything about the underlying mechanism ("Zero knowledge"). Or it can incorporate some external functional information such as which regions of the genome to consider, as in candidate gene studies, or by weighting interactions according to external knowledge, as in protein networks ("Mapping with knowledge"). In the presence of complete models of either an individual or a subsystem, where phenotypic consequences are predictable, much more complex genotype to phenotype functions would be possible ("Model-based mapping").



N3 *Gene/Transcription Regulatory Networks* - Nodes are labelled with gene transcripts and in some cases known transcription factors. Edges are representative of transcription regulatory mechanisms for example transcription factor binding.

N4 *Metabolic Pathways* - Nodes are labelled with metabolites and edges represent chemical reactions. Since many reactions require catalysis by additional molecules (such as enzymes), edges may involve more than two nodes thereby creating a hypergraph.

Notably, in each of these types, nodes generally represent the same type of biomolecule or at most feature two types of molecules (for example gene transcripts and transcription factors). With the increasing availability of different high-throughput data types, integrated biological networks are becoming increasingly reported, where nodes are labelled with any kind of biomolecule and edges again labelled with the processes by which they are related.

The strategies used to reconstruct biological networks vary according to the existing biological knowledge and data upon which they are founded. They can be classified into three categories accordingly:

1. **Theoretical Modelling:** This approach is based on existing biological knowledge (perhaps obtained via literature searches) and chemical/ physical laws (such as the law of mass action). No data in its raw form is used. This approach is successful for dynamic modelling of signalling pathways, transcriptional regulatory networks and metabolic pathways. It can also be used to make predictions of protein-protein interactions on the basis of the properties of the proteins rather than observations of physical interactions.
2. **Physical Interaction Modelling:** The nodes and edges of these networks are defined by the data. Edges are inserted between two nodes if they are observed to physically interact. This approach is usually only used to reconstruct static protein-protein interaction networks.
3. **Statistical Modelling:** This approach uses observations of data at the nodes of a network to infer edges. There are a range of statistical techniques which can be used to infer networks at a single snap shot in time or dynamic networks over a range of time points. This approach can be effective for both small and large data sets. In addition, statistical modelling can also be used in conjunction with theoretical models to provide a more detailed description of a system. For example, data could be used to infer rate parameters of a metabolic reaction.

We continue to describe the concepts and techniques used in theoretical and statistical modelling since these are used for the interpretation of global variation data sets.

3.2.2 Theoretical Network Models

The major cellular networks (N1-N4) are physical-chemical systems that can be characterized to a high degree based on a physical description collated from biological knowledge. Theoretical modelling strategies make substantial reductions to summarise these descriptions. Static protein interaction networks are usually reconstructed using physical interaction modelling approaches but the remaining three classes (signalling pathways, transcriptional regulatory networks and metabolic pathways) have a temporal component. Although they are more complicated to model theoretically and experimentally they can provide a more accurate representation of a true system.

Dynamic models can be classified according to three main criteria. Firstly, a model can be deterministic or stochastic. Systems involving very large number of molecules ($> 10^3$) are often modelled using deterministic models, while stochastic treatment becomes necessary for networks based on a smaller number of molecules. Secondly, a model can allow for spatial heterogeneity or it can assume spatial homogeneity. Spatial homogeneity is conceptually and computationally attractive but simple spatial models can be constructed by decomposing space into a few disjoint compartments. Thirdly, a model can be discrete or continuous in time and this affects the characterisation of the state space the models.

Metabolic pathways have by far the longest history of modelling and the classic models like Michaelis-Menten are almost a century old. Metabolic Pathways are almost always modelled deterministically by ordinary differential equations (ODEs). Modelling individual reactions is very challenging in itself and parametricising a single reaction is difficult [28]. Going from single reactions to a complete metabolic pathway is a difficult task and there are a series of approximate approaches which have been developed to reduce the complexity of the

problem. Savageau [112] in a series of papers starting in the late 60s used ODEs with terms only having powers of concentrations to analyse smaller pathways. Metabolic Control Analysis (MCA) was developed early 1970 by Heinrich, Small, Kacser and Burns. MCA uses linear approximations around an equilibrium point to explore the consequences of changes in substrates and enzymes. Flux Analysis has been pursued by a variety of authors and completely ignores the kinetics of reactions and only describes the metabolic capabilities of a network in terms of the stoichiometry of the underlying reactions. This is clearly a serious simplification but has been useful in describing the metabolism of different bacteria.

Signal Transduction Pathways can be modelled by both stochastic and deterministic models. Their overall dynamic behaviour still have not been clarified and especially the overlap between different networks seem enigmatic.

Regulatory Networks are the focus of much current research. The earliest models goes back to the early 1960 just after the operon model had been proposed by Jacob *et al.* [64]. A variety of models were proposed including systems of ODEs and Boolean networks, but progress was hampered by lack of appropriate data. Expression data, although noisy, has clearly induced an insurgence in modelling and inference. One major innovation has been stochastic modelling. Regulatory networks which involve only a few molecules can be modelled stochastically [139].

3.2.3 Statistical Network Models

In instances when edges are not explicitly defined by data, often statistical models are used to infer them. Networks inferred using statistical modelling aim to provide insight to biological network characterised by biomolecules and their interactions, but the nodes and edges of a network inferred statistically have different labels. Nodes are considered random variables (e.g. the concentration of a molecule could be considered as a continuous random variable where as a genotype or CNV could be considered a discrete random variable) and edges reflect dependence between nodes rather than functional biological relationships.

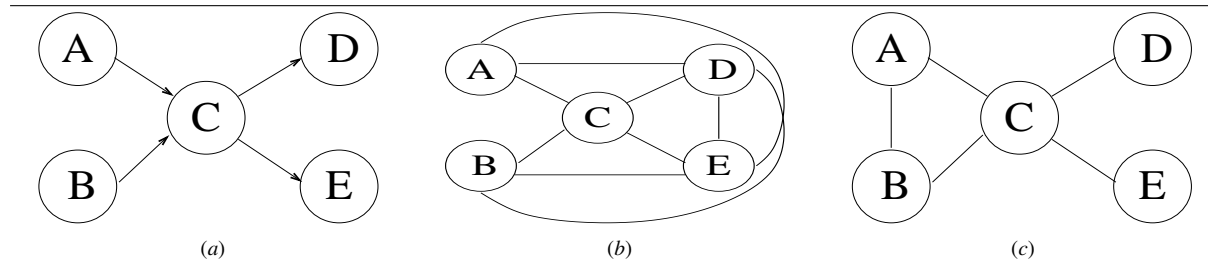
There are several important statistical principles and frameworks which are commonly used to infer statistical networks from biological data. We describe them below:

1. *Correlation*: Intuitively, two quantities or variables are independent if knowledge of one variable tells you nothing about the other, and they are often said to be correlated if they are not independent. Correlation can be estimated using a correlation coefficient which takes values in the range $(-1,1)$; it quantifies the strength and direction of the linear relationship between two variables. A quantity X is perfectly correlated with itself, hence the pair (X, X) has correlation coefficient 1 and $(-X, X)$ has correlation coefficient -1. Correlation coefficients between biological quantities can be estimated from observed data and presented in a correlation matrix where the entry in the i th row and j th column is the correlation between quantity i and quantity j . Correlation matrices can be used to construct undirected networks (e.g. co-expression networks), either by placing an edge between two nodes if their estimated correlation exceeds a threshold or by using a soft mapping. The latter approach weights edges according to the strength of correlation.
2. *Conditional Independence and Partial Correlation*: Correlations are a very naive way of looking at relationships between variables and can be misleading particularly because they do not reflect non-linear relationships and many correlations can be observed by chance or confounded by other variables. A more reliable measure for assessing evidence of dependence between variables is partial correlation. We demonstrate these concepts with the following example. Suppose the expression of gene D and E are correlated but only because they are regulated by the same transcription factor C (figure 4 (a)). Given the activity level of this transcription factor, the expression of gene D and E are independent. In this instance, genes D and E are said to be conditionally independent given C or equivalently, the partial correlation of D and E given C is zero. More generally, C could consist of multiple variables giving rise to higher order partial correlations.
3. *Dependence Graphs*: Complex dependence structures between variables can be represented by dependence graphs. They are undirected and can be constructed on the basis of partial correlations between variables. An edge is placed between two nodes in the graph if their partial correlation given all the other nodes in the graph is non-zero. Estimated correlations and partial correlations from data can be used to construct a graph and dependence structures can be easily extracted from the graph. Correlation and partial correlation alone cannot be used to infer causality and directed networks, particularly when inferred from static biological

data from the same source. For example gene networks constructed from static gene expressed data are usually undirected. Without additional information such as known transcription factors, genetic data or additional time series data, it is rare that statements are made about the direction of relationships. However in some circumstances it is possible to define a set of directed graphs which are consistent with the observed dependence structures (see below and figure 6).

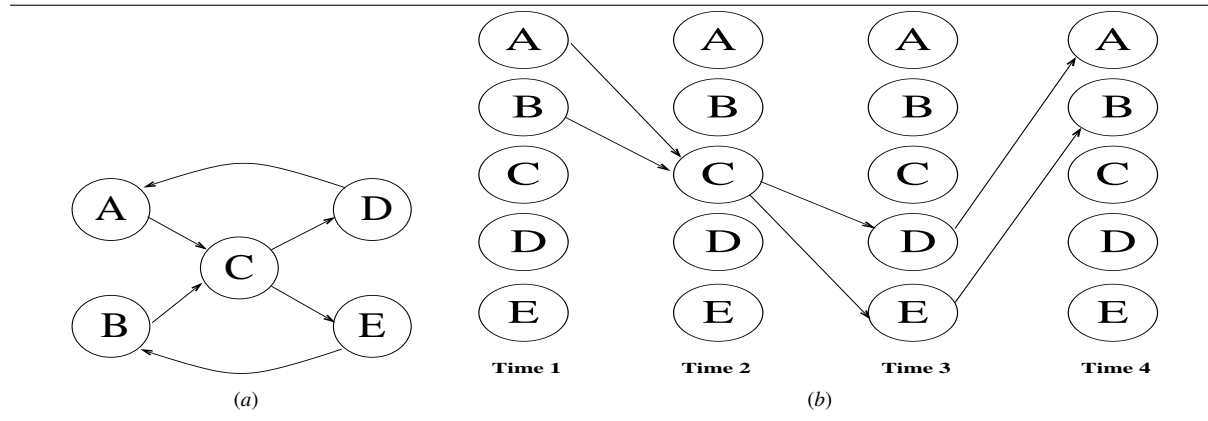
4. *Directed Graphs and Causality*: Adding directions to network edges provides a way of representing causality. A directed edge from node A to node B represents the fact that A influences B. For example SNP A might affect the expression of gene B. Directed edges can sometimes be inferred between nodes from different sources of data or by exploiting the flow of information which starts at DNA [143], alternatively, direction and causality can be inferred with the temporal data at a suitable resolution. Without temporal or other information, a set of directed graphs can also be derived which are consistent with the dependence graph. For large networks there may be many of these thereby making validation and interpretation difficult. Directed Acyclic Graphs (DAGS) are directed graphs which do not contain any loops. This additional property makes them particularly efficient and easy to work with.
5. *Dynamic Bayesian Networks*: Biological systems often contain feedback loops and complete cycles which are forbidden in DAGs. However, these can be modelled using Dynamic Bayesian Networks; They are a special type of DAG with a series of temporal levels where each node features at each time point in the network. Edges are directed forwards in time enabling feedback loops and cycles inherent in the biological systems to be incorporated without creating direct cycles in the graph. Hence, the efficient model fitting of DAGs can be exploited and to infer networks which can be more readily interpreted biologically.

Figure 4 (a) A network which represents a simple gene regulatory mechanism without feedback loops. Each node represents a gene; Genes A and B directly regulate gene C which simultaneously regulates genes D and E. (b) A co-expression graph which might be inferred from static transcript abundance data observed from system (a) on the basis of estimated pairwise correlations. (c) A dependence graph which might be inferred from static transcript abundance data observed from system (a) on the basis of partial correlation. Notice that the topology of this network closely resembles the topology of (a).



These principles are illustrated by figures 4 and 5. In figure 4 the distinction between a correlation based network and a partial correlation based network is illustrated with reference to a simple acyclic gene regulatory mechanism. Notice that the correlation network is nearly complete (with only the link between gene A and B missing); if there is a way of getting from one node to another following the directed arrows (possibly via another node) then these variables will appear correlated. Correlation based networks might be able to identify a module of co-regulated genes amongst a larger set but do not in general reflect the nature of the dependence structure. A dependence graph provides more information regarding structure. In particular the topology of the dependence graph and the mechanism differ only by one edge and hence can be more readily interpreted biologically (see the following subsection). The additional edge between gene A and B is present since the activity of gene A is not conditionally independent of B given C. For an extreme example suppose $C = A + B$, then given C and B, A is completely determined hence cannot be conditionally independent of B. For a more technical explanation of how dependence graphs can be extracted from directed graphs, see Lauritzen [80]. In figure 5 we illustrate a more complex regulatory mechanism involving the same set of genes with feedback loops. By replicating the nodes of the graph at multiple time points feedback loops can be modelled without creating directed cycles in the graph. This modelling strategy is however dependent upon the availability of temporal data.

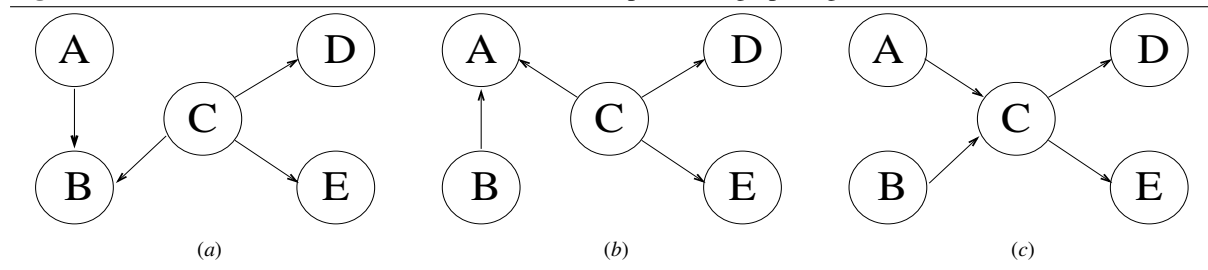
Figure 5 (a) A gene regulatory mechanism with a feedback loop. The system is as in figure 4 (a) with two additional feed back loops; Gene D regulates gene A and gene E regulates gene B (b) An example of how (a) can be represented with a directed acyclic graph and hence modelled using dynamic Bayesian networks



3.2.4 Biological Interpretation of Networks Inferred Statistically

Biological interpretation of an inferred network is paramount, but it is a difficult task, particularly with respect to causal interpretations. Biologically implausible relationships are usually excluded prior to network inference, but there may be multiple biological explanations consistent with an inferred statistical network. Further ambiguities can arise when several statistical networks explain biological data equally well. The ability to reconstruct an informative biological network from statistical data relies on the topology of the statistical network and the physical system to be similar, but plausible biological explanations must be explored before a biological network can be constructed. For example, the regulatory gene mechanism described in figure 4(a) could be represented by the dependence graph (figure 4(c)) but there are three possible directed acyclic graphs consistent with this dependence structure as illustrated by figure 6, each of which could have different causal and biological interpretations.

Figure 6 The set of DAGs which are consistent with the dependence graph (figure 4 (c))



Causal hypotheses can be explored by observing perturbations of a system. Comparing predictions with observations of the effects of a perturbation on a system might support or validate an existing hypothesis. There are obvious limitations to this approach in human systems but it is widely used in model organisms and cell lines reference section on model organisms. For example, models (a), (b) and (c) in figure 4 could be distinguished by systematic perturbations of genes A, B and C.

Edges in statistical networks represent dependence structures and possible hypotheses surrounding the etiology of disease phenotype. A biological network representing the same system might share a similar topology, but the edges would represent physical processes. For example, in figure 4(a), genes A and B might encode distinct transcription factors which bind upstream of gene C to initiate its transcription. Consideration should also be given to the interpretation of nodes in a graph. For example in many instances (and in figures 4 and 5) a node is often labelled with a gene and represented solely by mRNA transcript abundance however, biologically perhaps this node should be decomposed with genetic, transcript and protein components such that an integrated network can be constructed relating these units. Such a decomposition would require multiple data sources and computa-

tionally intensive statistical methodology but the resulting networks could potentially be more readily interpreted biologically and useful to suggest putative causal mechanisms.

3.3 Genealogical Relationships

Genomic variation can be observed at three levels:

1. Across cells within an individual: All cells in an individual are related by an ontogenic tree.
2. Across individuals within a population: All individuals in a population are related by a pedigree. In most cases, the pedigree relating individuals of a population is not of primary interest but rather how genomes are related. This information can be represented by an ancestral recombination graph (ARG).
3. Across species: Species are related by a phylogeny - at times with some loops due to hybridization or horizontal transfer.

Thus, the relationship between any set of cells can be traced through appropriate genealogical histories. Consider for example, a cell sampled from the nose tip of a mouse and a cell sampled from the toe of a human. These cells could be traced back to the zygote starting the individuals in 20-40 cell duplications. The individuals and their genomes can be traced back in pedigrees and ancestral graphs in 10^4 - 10^7 generations or 10^5 - 10^6 years and finally the species could be traced back in a phylogeny in 10^8 years or 10^7 - 10^9 generations.

Full genomic sequencing efforts initially targeted genomes at the species level, particularly those known with many genomic differences. Population genomic variation followed and is currently typically observed at a subset of markers in the genome rather than in its entirety. Finally there is genomic variation observed at a cellular level where relatively few differences are expected between cells sampled from the same individual. In the mouse-human example, there would roughly be a difference every 50 nucleotides between the species genomes, one difference in every 1000 nucleotides in two population genomes and possibly a difference every 10^7 - 10^8 nucleotides for cellular genomes.

Species and individual variation contribute rich source of information which can be of central use in integrative genomics. For example, population based studies can exploit the genealogy relating individuals and similarly, comparative studies can exploit differences between species. However, despite the fact that most current high throughput technologies make observations on cellular quantities, cellular relationships within an individual are rarely explicitly used.

The three categories of evolution described in this section provide different sources of information. Genomes change at the species level is ideal for measuring rates and selection. Genomes at the population level have correlations (linkage disequilibrium) along its sequences making genetic mapping possible. At present, genome differences at the cell level contributes less than the other categories. Expression data often comes from a collection of cells, but the variation in mRNA levels in those cell reflect differences in state rather than differences in genetic material. At other levels such as regulatory or protein networks, shape and behaviour etc, the evolutionary models are less well developed as the data available is much poorer than for genomes.

3.3.1 Genealogies relating cells in an individual

An individual consists of a large number of cells that has been created by a series of duplication that can be traced back to the zygote. For humans this is about 10^{14} cells corresponding to ≈ 45 generations. The relationship of all cells in an individual can thus be described by a traditional phylogeny. Traditionally cells of an individual have been viewed as genetically identical (with few exceptions such as cells involved in immune response, egg and sperm cells), but a cell and its duplicated offspring can differ in several heritable properties: Chromosome rearrangements, CG-methylation, microsatellite repeat numbers. Thus knowing the genomic sequence for all cells would allow the recovery of the cell phylogeny at high degree of certainty [44, 135]. Although for instance 50 events/cell replication could be a high estimate (for this purpose optimistic), the large number of cells descending from a given ancestor could guarantee that the phylogenetically necessary variation always was available. I.e. if somatic mutation rate is low, this could be compensated for by more sequencing. Only few studies have been

undertaken of cell phylogenies, but this will change. In contrast, cancers have been studied extensively due to the intense interest in this disease and fast chromosomal evolution in cancer cells.

3.3.2 Genealogies relating individuals of a population

A pedigree can be used to describe how individuals are related and is defined by assigning parents to individuals. The pedigree for a typical current individual is generally known 3-4 generations back in time. There are famous cases where pedigrees are defined for many more generations back in time like for instance the Icelandic pedigree [58] going 12 generations back and involving almost half a million individuals. How much pedigree information can be inferred given full knowledge of all genomes (if this was possible) remains an open question. The number of ancestors to an individual grows almost exponentially as a function of generations back in time, however the total amount genetic material ancestral to an individual remains constant. This is a result of recombination events (at least one per generation in the formation of germ cells) which breaks up genetic material into segments; the number of segments of ancestral material grows, but most ancestors further back in time do not contribute genetic material to a specific individual. Consequently, pedigree information far back in time will be of little genetic relevance.

The genealogical structure necessary to describe the history of genomes is the Ancestral Recombination Graph (ARG). This is a generalisation of the classical phylogeny that also traces recombination events, where a sequence has a mother and father sequence (although not gender labelled) carrying ancestral material to the left and the right of the recombination point. An ARG can in principle be embedded in a pedigree, since sequences are clearly found in individuals but in most instances, in this context, the individuals are ignored.

Treating individuals in a study as independent ignores the genetic dependence which is inherent between any set of individuals. Failing to account for this dependence can confound detection of genetic factors influencing phenotype [101].

Even in instances where relatedness of individuals is partially known, such as with family studies, it does not fully describe the ancestry of genetic material. This requires detailed modelling of both recombination and coalescent events. This is usually done using stochastic processes [69, 60] and defining a probability measure on genealogical structures.

3.3.3 Genealogies relating species

There are in excess of 10^7 known species. Genomes evolves by a series of biochemical events like substitutions, insertions, deletions, duplications, inversions, transpositions and translocations that occur with characteristic rates. The effective rates observed when comparing genomes from different species will then be increased or decreased if these events have an influence on fitness. There is certain time homogeneity of closely related species at three levels: 1) genomes will have a similar content and individual regions can be compared and 2) the basic rates and nature of evolution is also similar, so observing evolutionary change, which by definition will have been at another time point, can be treated as if it had happened in the species of interest and finally 3) phenotypes of species will also be similar making interpretation of genomic change realistic.

The main contributions of genome evolution can be considered in three categories: (1) the observation of basic rates of events, (2) the identification of regions under selection and (3) the functional interpretation of the actual content of the genome in terms of molecular mechanism. Basic rates of evolutionary events (such as mutation or recombination) are of intrinsic value in understanding the mechanism of organismal change. The strength and direction of selection are of great interest since it can be a consequence of genome function. Regions under negative (also termed purifying) selection evolve at a decreased rate relative to neutrality. They are of interest because they can be indicative of genome function, which has substantial effect on fitness. Conversely, regions under positive selection experience an increased rate of evolution relative to neutrality. Although they are less prevalent than negatively selected regions, they are also of great interest to study since they can be indicative of functional regions which adapt to environment.

3.4 Knowledge

Without exception, all studies of global biological variation exploit existing knowledge to some degree and yet defining what constitutes ‘knowledge’ is difficult; i.e. how do you know when you know something? Coupling knowledge with a measure of uncertainty provides one way of resolving this question although defining this measure of uncertainty is also non-trivial. The Bayesian framework provides the structure to coherently incorporate uncertainty about knowledge into a model by representing it using a probability distribution. This probability distribution can then be updated given observed data and where necessary averaged over to make statements about a quantity of interest.

In any given study, there is usually a set of things which are assumed to be true (with probability 1). These are usually experimentally tested and validated by multiple researchers independently and generally accepted by the scientific community to be facts. They vary according to the study but usually strongly influence study design and data collection (table 3), with some also providing the foundations for model construction and analysis.

DATA/MODEL	KNOWLEDGE
Genetic Markers	A set of informative loci which vary in a population, and where these markers are located in the genome
Gene Expression	Annotation of the genome with genes and transcripts which can be mapped to these genes
Exon Expression	Annotation of the genome with introns and exons of genes and transcripts which can be mapped to these locations
Protein-DNA Binding	Protein specific antibody or antigen. Map of entire genome for tiling array design.
Protein/metabolite abundances	Molecular composition of targets and their mass.

Table 3: Some examples where biological knowledge is used in study design and data collection

In particular, the other concepts described in this section are also founded on biological knowledge that is accepted to be true; the central dogma underpins the concept of a mapping from genotype to phenotype, knowledge of biomolecules which physically interact motivate development of network models, and knowledge about evolutionary processes motivates the use of genealogies. Furthermore as we go on to describe, knowledge that there are unobserved structures present in data (all types but particularly sequence data) motivates development of models to infer these unobserved states.

More specialised levels of knowledge are also used in many studies, but these are usually used to build a model or set of models to analyse data or are used in the interpretation of results. They might also be considered facts or assigned a degree of uncertainty (assigning priors in a Bayesian framework or weights). For example in a network model, there might be a core set of nodes and edges which are fixed according to experimentally validated interactions. In models of genotype to phenotype mapping there might be the assumptions made about how a genetic variant influences disease risk. Even if knowledge is not used in the model development, knowledge is almost always used to add biological context to statistical findings.

This level of knowledge is usually ascertained via literature documenting previous experiments and observations of biological systems. The increasing numbers of studies of biological variation, necessitates the development of a consistent representation of knowledge and tools such that it can be efficiently exchanged between researchers. There are several tools for cataloguing and collating knowledge (K1-4) and they are important to facilitate the refinement of modelling techniques and improve understanding of biological processes. Furthermore, since follow up studies are costly, it is important to effectively utilise existing knowledge to minimise the number of false positive findings which are further investigated.

- K1 Ontologies and Databases: Ontologies are structured databases which aim to provide a consensus vocabulary for knowledge exchange. The most famous example is Gene Ontology (GO) which aims to describe gene function. They are typically organised as a tree or DAG reflecting natural nested structure of terms [11, 7, 48]. Other databases also exist with differing levels of structure

- K2 Systems Biology markup languages: These provide a set of conventions for how to formulate models to be exchanged and allowing computer parsing. It is based on a more general set of conventions called eXtensible Markup Language (XML). First versions were publically available in 2003 and have since seen version 2.0 with extended flexibility. At present it is directed towards cellular biochemical models and does not place restrictions on the language in which the model is formulated (such as MatLab or MATHEMATICA).
- K3 Process Algebras: These are disciplines of formal computer science which are devoted to the description and analyses of processes. Examples include p-calculus, lambda calculus, PEPA and Petri Nets. Although they are founded on theory, process algebras have many biologically realistic properties such as allowing process interaction and refinement of descriptions. In particular, Calder *et al.* [15] and Phillips and Cardelli [97] use process algebras for automating ordinary differential equation models for simple biological systems. Potentially they could become a very powerful tool for systems biology although they currently lack flexibility to incorporate several continuous features necessary in biological models, such as space and concentration [16, 5, 98, 3, 77].
- K4 Text Mining Methods: These are automated techniques which extract information in a coarse manner from large bodies of text which could not be read by a single researcher. They are based on the simple underlying principle that words in articles can be tabulated according to frequency, contextuality, combinations, information content and more. This very efficiently allows linking genes with genes, biological objects with sets of properties and much more. The widespread use of text mining shows that in the large body of articles information is encoded in a way that is not detectable by simple sequential reading of a smaller set of articles. Text mining is progressing fast and with the increased use of ontologies and associated well-defined terms, text mining will increasingly acquire abilities closer to traditional reading [105].

3.5 Hidden Structures

It is very fundamental aspect of biology that one can observe certain quantities that are a function of something that cannot be observed. This is inherent in statistical modelling where there is often a hidden model that generates observables and we try to make statements about the hidden model. The last two decades have seen this taken a step further in the statistical model itself have been split into a hidden part that influence a part that generates the observables. It is a very general principle inherent in the way inference is done in biology. The most famous example is the class of hidden Markov models that now are ubiquitous and have very widespread applications in the biosciences. Other examples include stochastic context-free grammars and Kalman filters which also have biological applications (table 4).

There are two major classes of models:

1. **Hidden strings describing the generation of observed strings:** The most famous example is the Chomsky linguistic heirarchy [23] of grammars. A grammar is a finite set of rules which generate words or sentences in a language. They were originally deterministic and used to define which strings of letters belong to a language and which do not. If rules are assigned probabilities, then strings will be generated from the rules with different probabilities and the grammars are stochastic. There might be multiple ways to generate the same observed string; in this case, the probability of an observed string is the sum over all possible ways it can be generated. If there is only one series of rule applications, then it is simple to assign a probability. The hidden structure behind a string of letters is the series of rules used to generate it and this can be represented by another string of characters.

The bottom two layers in Chomsky's hierarchy of grammars i.e. regular and context-free are widely used in biology. Typically their stochastic versions (hidden Markov models (HMMs) and stochastic context free grammars (SCFGs) respectively) are used as hidden models for biological sequences (of DNA or RNA for example). HMMs have an enormous versatility. Major modelling areas in sequence analysis include gene structures, protein secondary structures, signals and alignment. Stochastic context-free grammars have especially been used to model RNA structures.

2. **Hidden Processes:** These have been used by for instance Lawrence, Rattray and colleagues, where a hidden Gaussian Process was used to model fluctuations in a transcription factor that influenced the activity of 5 gene products that could be observed. This was a very natural formulation that should be extended to more general settings and other cellular processes.

DOMAIN	OBSERVED	HIDDEN	REFERENCE
Haplotypes on a pedigree	Haplotypes	Inheritance	Lander and Green [78]
Isochores in Genomes	Sequences	Isochores	Churchill [24]
Sequence Alignment	Sequences	Alignment	Krogh <i>et al.</i> [74]
RNA Structure	Sequences	Structure	Sakakibara <i>et al.</i> [109]
Protein Secondary Structure (SS)	Sequences	SS	Goldman <i>et al.</i> [49]
SS relationships	Sequences and SS	SS Interactions	Abe and Mamitsuka [2]
Protein Genes	A Genome	Gene structure	Burge and Karlin [14]
Rate Variation	Homologous sequences	Fast/slow regions	Felsenstein and Churchill [40]
Hidden Dynamic Processes	RNA levels	Transcription factor	Lawrence <i>et al.</i> [81]

Table 4: Examples of where models of hidden structures are applied in biology; Their observables, hidden states and references.

Several of the above applications have natural comparative extensions. Goldman *et al.* [49] and Felsenstein and Churchill [40] are already comparative, while Burge and Karlin [14], Sakakibara *et al.* [109] was made comparative in Pedersen and Hein [96] and Knudsen and Hein [72] respectively. Churchill [24], Abe and Mamitsuka [2] and Lawrence *et al.* [81] could be combined with models of evolution to make comparative extensions with great benefit.

Modelling via hidden elements is very natural and will most likely grow tremendously in the future. The applications so far are very simple, where the hidden element is unambiguous. It is easy to imagine more complex models, such as more interacting cellular components or a large set of possible networks relating the observables.

4 Analyses

We group analyses according to types of data they incorporate starting from analyses of single sources of data (i.e. any one of G, T, P, M, E, F) ranging through increasingly integrative approaches which include data from non-singleton subsets of {G, T, P, M, E, F}. The data types coupled with concepts provide the foundation of analyses and table 5 summarises the analyses we discuss in this section according to the data types and concepts upon which they are founded. Since widespread genome-wide sources of data from P, M and E only recently became feasible for such studies, analysis with some subsets are rarely performed and we restrict our discussion to the most prominent analytical techniques.

ANALYSIS	DATA	CONCEPTS	OUTPUT
Genome Annotation	Full sequence data (G) Multiple Species	Biological Knowledge Hidden Structures Genealogies	Predicted annotation of genes, signals, structure and function.
Population Genetics	Genetic markers (G) across individuals	Biological Knowledge Genealogies	Linkage Disequilibrium Blocks, Population Structure, Inferred Ancestries, Rates of DNA Mutation and recombination, selection.
Clustering molecular features	Molecular abundances (T), (P), (M)	Minimal Knowledge	Modules of co-regulated/co-expressed genes or molecules
Molecular Network Analysis	Molecular abundances (T), (P), (M)	Biological Knowledge Networks	Networks of genes, metabolites or proteins
Association Mapping	Genetic marker data (G) Phenotype (F)	Biological Knowledge Genealogies G → F mapping	Genetic Markers or regions in genome which associate with a phenotype
Differential Analysis of Molecular Abundances	Molecular abundances (T) (P) (M) Phenotype (F)	Minimal Knowledge	Molecules present at differential levels under different conditions i.e. Biomarkers.
Molecular Quantitative Trait Analyses	Genetic markers (G) Molecular abundances (T), (P), (M)	Biological Knowledge G → F mapping	Genetic Markers or regions in the genome associated with molecular abundances
Regulatory Analysis	Genetic Markers (G) Molecular abundances (T) (P) (M)	Biological Knowledge, G → F mapping Networks	Directed networks of genes and molecules according to how they interact and are regulated.
Integrated Analyses (comparative based)	Genetic markers (G) Molecular abundances (T), (P), (M) Phenotype (F)	Biological Knowledge G → F mapping	Overlapping genetic associations, Candidate Genes, Putative causal mechanisms
Integrated Analyses (simultaneous)	G, T, M, P, F	Biological Knowledge G → F mapping Networks, Hidden Structures	Integrated networks, putative causal mechanisms, candidate genes

Table 5: A summary of analyses we discuss; the data and concepts they use and their output

4.1 Analysis of single sources of data

4.1.1 Species Level Genomic Variation Data; (G)

The genome of any (metazoan) organism is highly structured and encodes protein coding genes, RNA secondary structure and many regulatory signals. Predicting which nucleotides of the genome encode these features is termed

genome annotation and provides a crucial stepping stone towards understanding an organism. While some annotation is the result of functional experiments involving specific molecules, an increasing source of annotation is derived by making comparisons within and between genomes of a variety of species.

The analysis of the genomic sequence data of a single organism was introduced when sequencing first appeared. Protein genes, RNA genes and signals have distinct characteristics which enable them to be studied in a single genome. More specifically, the triple periodicity of genetic code, base pairing in RNA and physical characteristics of promoters provide the basis of many predictive models. The first successful human genome gene finder was developed in 1997 [14] with no reference to other genomes (since there were no other eukaryotes sequenced).

The simultaneous analysis of genomes from multiple organisms (comparative genomics) has emerged over the past decade and the domain of its application is increasingly expanding due to the large number of sequenced genomes [99]. The observation of conserved regions and other common features in the genomes can be used to make statements and predictions about selection, gene structure, RNA secondary structure and signals. Of these, selection is notable since the observation of different kinds of selection contributes tremendously to the “needle in haystack problem” of finding functionally important regions.

Many of the models developed to predict nucleotide function (for both single and multiple genome analyses) exploit two tier stochastic modelling: (1) a model of the distribution of the structure in question (e.g. gene structure) without knowledge of the nucleotide sequences, and (2) a model of sequence evolution conditional on the structure. For instance, a nucleotide will typically evolve slower in a coding region than in a non-coding region. Such techniques have successfully been applied with mammalian genomes, to annotate protein coding genes [14] and RNA secondary structure (e.g SCFGs, [72]). The two tier system can also be applied to annotate regulatory motifs. However, due to their short length, faster rate of insertion/deletion and the lack of simple characterising features it remains a challenge [110] and a subject of intense research.

Out of the main features encoded in the genome (protein coding genes, RNA structure, regulatory signals), protein coding genes are the most comprehensively annotated for the human genome. At present there are approximately 24,500 identified human genes, a large proportion of which are alternatively spliced. The discovery of RNA genes (i.e. those which do not code for protein) has been one of the greatest surprises of the last decade, firstly in their abundance and secondly with their large number of unexpected functions. The exact number and exact functions have been much harder to ascertain than the analogous problem for protein coding genes. Discovery and annotation of regulatory signals is more challenging than both RNA and protein gene annotation and consequently, they are the least well annotated of the three features.

Functional annotation is supplementary to that of protein, RNA and signal ‘status’ and is not easily defined without molecular biology and experiments. However, comparative genomics can be used to predict function on the basis of homology. This is founded on the idea that function can be inferred from known functions in similar organisms. Ontologies of gene function, although difficult to define, are very useful in creating a common vocabulary for describing genes and biological processes. Furthermore, large international efforts such as HapMap (<http://www.hapmap.org/>) and OMIM (<http://www.ncbi.nlm.nih.gov/omim/>) also catalogue variation at positions in the genome.

4.1.2 Human Genetic Variation Data; (G)

Human genetic data as described in Section 2.1, especially SNPs, have been studied to elucidate features of genetic variability. Key to much of the analysis has been the development of coalescent theory [69], which statistically models the evolutionary history of a sample of individuals described in Section 3.3.2. This model can incorporate the biological phenomena of mutation, recombination, population demography/structure, and selection, in principle allowing for estimation of these features. For example, the recent publications of completely sequenced human diploid genomes suggests two random chromosomes have $\approx 3.2 - 3.3$ million SNP mutations between them [82, 138]. The “standard” coalescent model, which assumes randomly-mating individuals and no selection, tells us we therefore might expect to see $\approx 3.2 \times \log_e 6e^9 = 72$ million locations of the 3 billion basepair genome sequence to be SNPs in a population of 6 billion humans.

Due to recombination, the expected linkage disequilibrium between SNPs in a sample of individuals (as measured by the correlation coefficient) decreases exponentially as distance increases, and higher-than-average levels

of linkage disequilibrium within human populations typically do not extend beyond 1-2Mb [102]. While recombination events are relatively rare, occurring on average once per chromosome per generation, coalescent-based techniques have been used to show that recombination activity appears to cluster into narrow 1-2kb regions of the genome known as “hotspots,” such that 80% of genome-wide recombination activity occurs in about 10-20% of the sequence [91].

It is also clear that genetic data varies detectably amongst human populations from different geographic regions. In particular the genetic diversity of populations decreases as their geographic distance from Africa increases, a pattern strongly supportive of a hypothesis that modern humans arose in Africa and spread across the rest of the world in a series of migration events with accompanying “bottlenecks” that resulted in loss of diversity [65]. Differences in genetic variation can even be detected between geographically close populations such as those in Europe, for which a recent study has shown that an individual’s DNA can be used to predict their geographic origin to within a few hundred kilometers [93]. Coalescent theory can in principle be used to estimate features of migratory patterns, though the number of possible scenarios to explore makes this difficult in practice.

Another recent finding has been the discovery of genomic regions showing high levels of conservation that might implicate selection. Signals of conserved DNA have been found at several loci in human populations, including the *lactase* gene and other loci thought to be involved in disease [92]. Comparisons between the three human populations of *HapMap* found evidence of selection in a dozen or so genes, including those involved in morphological characteristics such as skin pigmentation [133].

4.1.3 Molecular Quantities; (T), (M), (P)

It is rare that molecular quantities are observed without reference to some phenotype or specific context, however, the data can be analysed without using context or phenotype information. Structure and correlation within molecular observations of protein, metabolite or mRNA levels can be informative about co-regulation and genetic interactions. Clustering observations according to similarities they exhibit across samples is a very basic way. Clusters can be analysed for enriched functional themes using knowledge in existing data bases. Alternatively statistical networks can be inferred from the data and interpreted biologically to suggest possible modules of genes or molecules involved in common functionality.

In table 6 we summarise methods for the construction of gene networks from transcript expression data with some references. Although protein and metabolic networks can also be inferred in a similar way [136], it must be noted that these networks are distinct from protein interaction networks and metabolic reaction networks. The edges in the latter are inserted explicitly according to interaction data or by biochemical experiments rather than with abundance data.

NETWORK TYPE	DATA TYPES	INTERPRETATION	METHODS	REFERENCES
Co-expression Modules	Expression Sequence/Genomic	Groups of genes co-regulated	Clustering, Motif discovery and enrichment	Tavazoie <i>et al.</i> [125]
Co-expression Networks	Expression	Nodes represent genes, edges are present (possibly weighted) between genes if they are significantly co-expressed	Pairwise empirical correlations, adjacency functions to weight edges	Zhang and Horvath [141]
Regulatory Modules	Expression Known regulators	Nodes represent genes, edges are present between genes and their regulators if there is evidence of statistical dependence	Classification and regression trees (CART), Graphical models	Bonneau <i>et al.</i> [13], Segal <i>et al.</i> [117], Ernst <i>et al.</i> [38]
Probabilistic Regulatory Networks	Expression	Nodes represent genes, edges are present if genes are statistically dependent. Not necessarily representative of direct physical interactions.	Gaussian Graphical Models, Bayesian Networks	Schafer and Strimmer [116], Friedman [43]
Transcriptional Networks	Expression Sequence, chIP-chip (protein binding)	Nodes represent transcription factors and genes. Edges represent statistical dependence and/or characterised dynamic physical interactions	Differential Equations, Dynamic Modelling, Gaussian Processes, Hidden Markov Models	Ernst <i>et al.</i> [38], Sanguinetti <i>et al.</i> [111]

Table 6: Summary of Gene Network inference methods

Clustering and identification of gene modules has been particularly successful for the understanding of genes involved in various cellular processes particularly when expression is monitored over a time series [36].

4.2 Analysis of phenotype with another source of data

4.2.1 Analysis of phenotype with genetic data; (G+F)

Detecting associations between genetic variants and disease susceptibility or variation in a clinical phenotypic trait (termed disease/phenotype *association mapping*) has been a widely popular field for a number of years. It is founded on the assumption that there is a map from genotype to phenotype (section 3.1).

The techniques used to detect genetic variants associated with phenotype depend on study design and the relatedness of individuals selected for the study. There are two main types of studies:

1. Linkage mapping studies of families (i.e. *pedigree data*): Genetic markers and clinical phenotypes are collected from families of closely related individuals. The largest example is the Icelandic pedigree, consisting of genealogical information on 300,000 extant individuals (and 400,000 of their ancestors), with SNP information on 15,000 extant individuals with various diseases at 370K sites.
2. Genome-wide-Association Studies (GWAS) of distantly related individuals (i.e. *population data*): Genetic markers and clinical phenotypes are collected from individuals sampled from a population. An example is the Wellcome Trust Case Control Consortium, which maintains 550K SNP data from 14,000 cases of 7 major diseases and 3,000 healthy controls (<http://www.wtccc.org.uk/>).

Figure 7 illustrates the differences in the data resulting from these two sources. Techniques for analysing both sources of data often make use of hidden structures (section 3.5) and genealogical relationships between individuals (section 3.3.2).

Linkage mapping is useful for identifying broad regions (up to 10cM) harbouring phenotype-influencing location(s); such regions may contain many genes. While microsatellites were primarily used at the start of linkage mapping studies, the density and large-scale availability of SNP data has resulted in SNPs being the current marker of choice. Often, the \log_{10} Odds a marker is linked versus unlinked to a disease-influencing location (i.e. the LOD score) is reported as a score for determining genetic associations (see Dawn and Barrett [30] for a more detailed review). Linkage mapping techniques are most powered to detect highly penetrant single genes influencing a phenotype and have poor power to detect associations to complex diseases with multiple genetic components. However, such techniques have been able to identify genetic regions associated with Cystic Fibrosis [68], Huntington's disease, breast cancer, and muscular dystrophy. Some widely-used linkage mapping techniques include those described in Lander and Green [78] and Olson *et al.* [94].

GWA methods potentially allow for much greater resolution in the fine-mapping of phenotype-influencing loci. SNPs are by far the most common employed markers in GWA studies, though micro-satellites and, more recently, CNVs have been used to test for associations using similar methods. The most widely reported tests of association use single-marker approaches, where each genotyped SNP is tested for association with the disease phenotype independently of all other SNPs in a region, often reporting p-values based on chi-squared or other statistics as scores for testing associations (see table 2 in section 3.1 for examples of common mapping functions). However, methods that make use of linkage disequilibrium patterns between SNPs, i.e. haplotype-based approaches, have been successful in finding variants not detected by single-SNP techniques. Power to detect true associations via GWA methods depend primarily on sample size and the effect sizes and minor allele frequencies of the loci involved. Variants with small minor allele frequencies and/or small effect sizes require much larger studies in order to be detected. Recent studies have reported significant signals of association for SNPs with minor allele frequencies >0.05 and relative risks typically in the range of 1.2 to 2.0 [126]. However, in nearly all cases the location, minor allele frequency and relative risks of the true, presumably untyped, underlying causative variants is unknown. Interpretation of findings is made more complicated by the fact that some of the most significant association signals are often in "gene deserts."

4.2.2 Analysis of phenotype with molecular data; (F + T), (F + P), (F + M)

These analyses ((F + T), (F + P), (F + M)) all aim to distinguish molecular features which correlate with a single phenotype. They can be termed biomarkers or signature molecules since they can serve as an indicator of normal biological processes, pathogenic processes or pharmacological responses to therapeutic intervention. In the following text we outline the main analysis techniques and applications for each of the molecular phenotypes T, P and M with a phenotype.

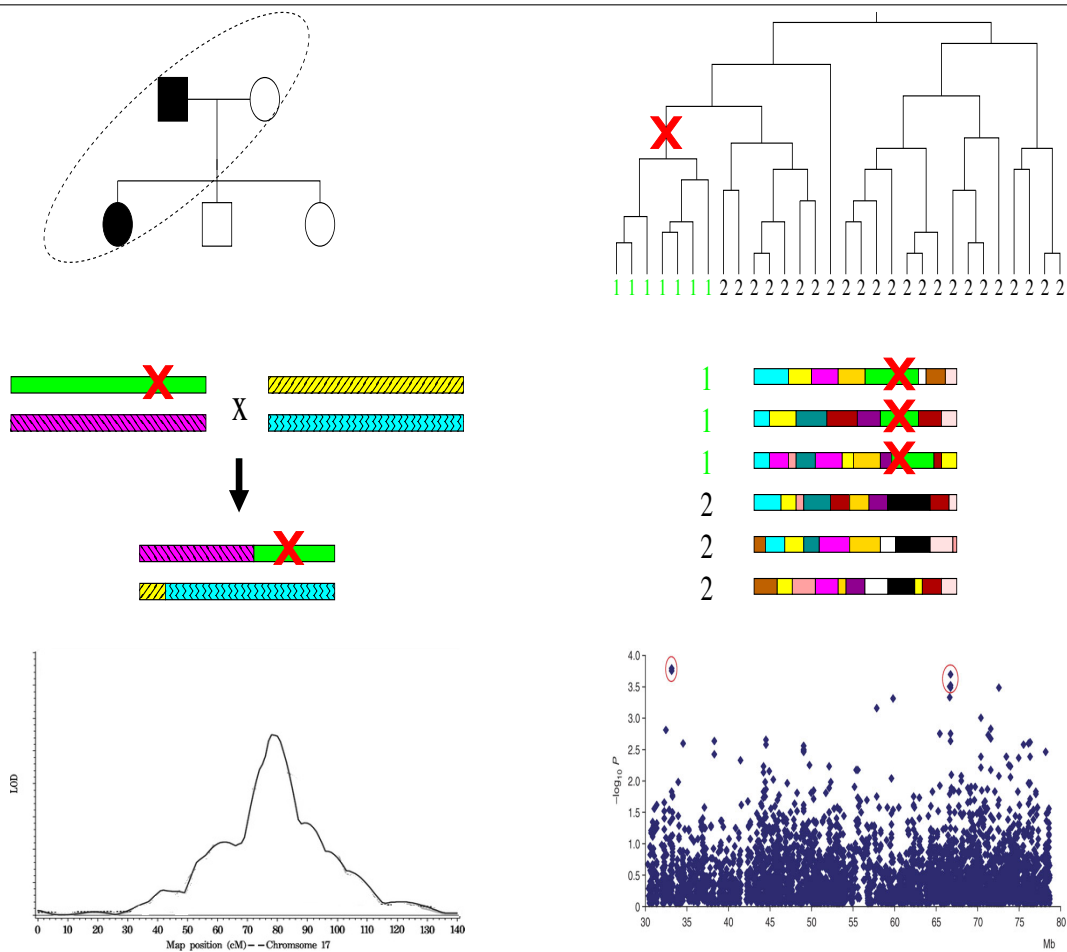
Differential Gene Expression (T+F)

One of the primary uses of gene expression data is to identify genes which are differentially expressed under different conditions. This covers a wide range of studies such as healthy versus diseased, control versus treatment, tissue comparisons and temporal changes in expression of homogeneous tissue. Determining which genes are significantly differentially expressed can be accomplished using a variety of scoring techniques.

Numerous disease studies over the past decade have exploited gene expression interrogation and analysis in applications for disease classification [50, 120, 137], prognosis [131] and drug response prediction [19]. A critical comparison and evaluation of microarray platforms [84] concluded that reproducibility and quality of gene expression data is sufficient for findings to be used in clinical environments. In particular, gene expression profiling is now exploited commercially in the clinical environment for classification, prognosis and treatment of breast cancer tumours [104, 103].

Similar studies are also emerging which detect the presence of alternatively spliced transcripts and isoforms which show significant differences between groups. The problem is statistically more challenging than detecting differential gene expression since there are many different types of splicing events which require different detection methods, furthermore, exon array probe sets typically contain fewer probes thereby increasing necessity of robust but sensitive probe filters and summary statistics. The most common way to assess whether a gene is being alternatively spliced is to use indexing methods [26] which involve computing a ratio of the abundances of different

Figure 7 Conceptual illustration of linkage mapping (left column) and genetic association studies (right column). **Top left:** illustration of a hypothetical three-offspring nuclear family (square = male, circle = female) consisting of one diseased parent and one diseased offspring (coloured in black). Each individual has two haplotypes representing any given genetic region: one inherited from the individual’s father and the other inherited from the individual’s mother. **Middle left:** illustration of both haplotypes in a chromosome region harbouring a disease-influencing genetic variant (red “X”) for the diseased parent (green solid, purple dashed slanting left), the undiseased parent (yellow dashed slanting right, blue wavy line), and the diseased offspring (bottom) circled in the top left picture. The offspring individual inherits regions from all four of her parents’ haplotypes in this example, including the region harbouring the disease-causing variant. **Bottom left:** LOD scores for prostate cancer association across a genetic region on chromosome 17 in humans using a linkage mapping approach applied to pedigree data of 147 families described in [79]. **Top right:** illustration of hypothetical evolutionary history for a set of present-day haplotypes (bottom) back to a single common ancestor many generations in the past (top), for a genetic region harbouring a disease-influencing genetic variant (red “X”). (This region is coloured green or black in the picture of the middle of the right column.) Connected lines illustrate related haplotypes, with the vertical length representing how far back in time until the two connected groups share a common ancestor. All haplotypes with a “1” show the disease phenotype; all haplotypes with a “2” do not. **Middle right:** Three present-day “1” haplotypes and three present-day “2” haplotypes. Note how colours switch more often in these haplotypes compared to the haplotypes of the family depicted in the middle left figure. This illustrates how the many recombination events in the extensive evolutionary history of these haplotypes act to break down associations amongst genetic variants, so that disease locations (such as the red “X”) can be better localised compared to using pedigree data. **Bottom right:** $\log_{10}p$ -value scores for prostate cancer association across a genetic region on chromosome 17 from a genome-wide association study approach applied to 12,791 unrelated Icelandic individuals described in [54]. Note how the score peaks are considerably more narrow compared to the linkage analysis results in the bottom left figure.



regions/exons throughout the gene relative to a robust measure of the overall gene expression. Different indexes can detect different types of splicing events, however complex events cannot always be determined in this way.

Alternative splicing is associated and indeed causal to several diseases [134] but with the recent development of a global exon array, increasing numbers of studies seek to find differences in alternative splicing patterns which correlate with disease phenotypes [129, 119].

Clearly, gene expression profiling has provided many insightful and useful findings over the past decade but, it has limitations. Transcriptional regulatory processes are highly context dependent and sensitive to environmental stimuli, so findings relating to a specific disease or process are dependent on the origin of the sample of cells interrogated and whether the disease related genes or processes are active in these cells. A recent Icelandic family study (Emilsson *et al.* [37]) comparing gene expression profiles harvested from blood and fat tissue demonstrates the sensitivity of gene expression to tissue type: the authors found over 70% of genes expressed in fat to be significantly correlated with Body Mass Index compared to only 9.2% of genes expressed in blood. It is not surprising therefore, that many of the successful gene expression studies of the past decade feature cancer, where tumour cells can be sampled for expression profiling. For other complex diseases (e.g. schizophrenia, asthma), there may be no obvious diseased tissue that can be easily sampled.

Functional biological interpretation of differentially expressed genes is also limited; causal and reactive gene expression traits cannot be distinguished without additional data (for example genetic, temporal) and furthermore, there is considerable debate over how well transcript abundance correlates to the actual amount of protein being manufactured by a gene, which is presumably the feature of the central dogma most likely to influence the clinical phenotype. This is discussed further in section 4.3.

Differential Protein Abundance/Structure (P+F)

In a similar way, protein abundances can be analysed under a range of conditions to identify proteins which are present at significantly different abundances under different conditions. These proteins (or indeed any other substance which can be used to differentiate between different conditions/phenotypic states) are termed biomarkers and as with gene expression due to the tissue specificity of protein levels, the majority of diseases for which proteomics has successfully identified biomarkers are those for which disease tissue or serum can be readily sampled. Examples include cancer [57], heart disease [86], autoimmune diseases [140] and Alzheimer's [142].

Some techniques used to quantify and characterise proteins (NMR for example) can also distinguish between various different states of proteins. In particular phosphorylated proteins can be identified and patterns of phosphorylation compared between normal and disease tissue. In addition to identifying these "biomarkers" for disease status, investigating and inferring the functionality of proteins, particularly those which vary in structure and/or abundance in disease states can also be facilitated by the collection of proteomic data.

Differential Metabolite Abundance (M+F)

As with protein and gene expression, metabolites (and the processes from which they are derived) are context sensitive but since metabolic experiments and quantification can be done with a range of samples from single cells to serum or tissue, this can be highly informative.

Perturbations of cellular processes due to disease states can be identified and characterised by metabolic profiling, such that different diseases, subtypes of disease or response to drugs can be identified according to the levels and/or presence of specific metabolites which thereby constitute an associated metabolic signature or fingerprint [107, 39, 108] Furthermore, the composition of metabolic signatures (usually ≈ 10 distinct metabolites) can provide information about the processes perturbed by disease according to their biochemical properties and existing knowledge about the reactions from which they are synthesised or converted.

Like proteomic profiling, metabolomic profiling is not yet widely employed for discovery driven research although offers great potential for pathophysiology. However, its success and wide deployment is also dependent upon appropriate and available cell/serum/tissue relevant to the disease/process being studied together with the ability to accurately collect and interpret metabolic data using statistical models. There are additional applications to drug discovery since many drugs and drug responses also have metabolic signatures [67].

4.2.3 Analysis of genetic data with molecular data; (G + T), (G + P), (G + M)

A natural way of combining information from genetic (G) and molecular data (T, P or M) is to use abundance levels as a quantitative phenotype and assess whether they are significantly associated with genetic variation using the techniques described in section 4.2.1. Considerable work has been done in this area for transcript abundance levels, i.e. gene expression, whereas studies with metabolite and protein abundances are only just beginning to be reported. Regions of the genome or individual genetic markers which are found to be significantly associated to a specific transcript, metabolite or protein are termed expression, metabolite or protein trait loci respectively (eQTL, mQTL and pQTL). They can be categorised according to where the markers lie with respect to their target; if they reside on the same chromosome or within a window around the gene they are cis-acting, otherwise they are trans-acting.

eQTL Studies

The combined study of gene expression traits and genetic markers in a segregating population was initially introduced by Jansen and Nap [66]. Following eQTL studies in yeast, Cheung *et al.* [21] and Schadt *et al.* [114] motivated similar studies in human with evidence of familial aggregation of gene expression traits in human cell lines, and heritable gene expression traits.

Both genome-wide linkage [90, 89] and genome-wide association studies (Cheung *et al.* [22], Stranger *et al.* [122]) have been used to search for eQTLs. The results of such studies are not entirely concordant, but as discussed by de Koning and Haley [31] and Pastinen *et al.* [95] this is not surprising given the differences in platforms, study designs, sample sizes, and analyses. However, common to these studies is an enrichment of cis-acting eQTLs amongst the eQTLs with the most significant effects, with the most significant cis-acting eQTLs being able to explain over 50% of the variation of its target expression traits. Interpreting eQTL effects must be cautioned since large effects in particular can be the result of confounding; a cis-acting eQTL in strong linkage disequilibrium with a SNP in a coding region of its target gene can affect the binding affinity of the mRNA rather than transcript abundances directly.

More recently Dixon *et al.* [32], Goring *et al.* [51], Emilsson *et al.* [37] and Schadt *et al.* [113] each investigate the genetics of global gene expression. Though differences in study design again impede direct comparison (notably the different samples sizes, data type, and platforms and cell types used to get the expression data), the most heritable traits in general have the strongest cis-acting effects in each of the studies. Despite the fact that some cis-acting eQTLs can have large effects, on average they only account for a small proportion of heritability of their target traits. The average heritability of clinical phenotypes is generally reported to be $\approx 25\%$ where as Dixon *et al.* [32] for example, report on average the top associated SNP to an expression trait can explain only 18.2% of heritability. This suggests other trans-acting effects or non-additive genetic effects contribute to their heritability. However these studies find little evidence of strong trans-acting eQTLs which may be reflective of insufficient power to detect multiple small trans-acting genetic effects.

The studies by Emilsson *et al.* [37], Schadt *et al.* [113], Dixon *et al.* [32] and Goring *et al.* [51] also highlight tissue specific differences in genetic regulation of gene expression. They target gene expression in blood, fat, liver, lymphoblastoid cell lines and fresh blood lymphocytes respectively. In particular, Emilsson *et al.* [37] make direct within individual comparisons between gene expression in blood and in fat. The choice of tissue can affect the number and size genetic regulatory effects and as we discuss in section 4.4.1 the interpretation of findings with respect to another phenotype.

At a more detailed level, Kwan *et al.* [76] use exon tiling arrays to investigate the effect of SNPs on gene expression, thereby allowing the detection of genetic variants associated with the expression of specific exons which might also be alternatively spliced. They detect a total of 324 genes out of 17,897 which show exon specific expression association with one or more cis-acting SNPs. Their results are categorised according to the type of splicing event and suggest that genetic regulation of gene expression is more complex than previously shown; only 39% of detected genes exhibit expression association with a nearby SNP across all exons, 29% show association of transcription initialisation or termination and the remaining genes are alternatively spliced e.g. via exon skipping.

The effects of other forms of genetic variation on molecular abundances can be studied in a similar way, in particular, there is recent work associating regions of copy number variation with transcript abundance. Contrary to SNPs which may be found in gene deserts, regions of copy number variation typically contain one or more genes

and thus might be expected to directly affect transcript abundances and proteomic expression of these and other genes. Stranger *et al.* [123] suggest that the relative regulatory contributions of CNVs to SNPs is small (17.7% : 83.6%) with little overlap. Furthermore, they find that target transcripts of most significantly associated CNVs do not lie in the region of copy number variation which suggests the effects of CNVs extend to the disruption of other regulatory mechanisms. As technology and ability to accurately genotype and define CNVs improves, the observed contribution of CNVs is likely to increase. At present, power to detect associations with CNVs is impaired by the difficulties in genotyping and defining regions of CNVs compared to identifying SNPs.

mQTL and pQTL Studies

The effect of genetic SNP variation on global metabolite and protein abundance is less well researched, although with these data types becoming increasingly affordable and practical it is likely that more studies of this nature will follow, especially since these studies could provide additional functional insight at a molecular level. Metabolites and proteins are more directly related to processes affecting phenotype but also exhibit more dynamic behaviour than transcripts and are highly sensitive to other non-genetic factors thereby making it more difficult to find genome-wide significant associations.

Notably, there are two recent studies Gieger *et al.* [47] and Melzer *et al.* [87] which are the first to report metabolite and protein quantitative trait loci respectively. Melzer *et al.* [87] show that human protein abundance levels (8/42) can exhibit strong association with SNP data with 7 out of the 8 strongest effects being *cis*-acting. The mechanism driving one of these effects (IL6) is well characterised and validated by independent studies to be the result of a SNP encoding an amino acid change. Two of the other associations are thought to be related to gene copy number variation although these are not validated.

In the first metabolome-wide association study [47] failed to find any mQTLs at genome-wide level of significance for metabolite abundances considered in isolation, but they report improve power by considering ratios of pairs of metabolites which are related to the substrates and products of an enzymatic conversion. Using this approach, they identified the FADS1 gene (among others) and deduce that the minor allele variant of the associated SNP reduces the efficiency of a fatty acid (*delta*-5 desaturase) reaction. This finding is also associated with cholesterol levels which we discuss further in section 4.4.1.

4.3 Analysis with multiple molecular data types; (T + P), (T + M), (M + P), (T + P + M)

The transcriptome, the proteome and metabolome are interactive, thereby motivating the simultaneous analysis of these data types. The simple flow of information from DNA to protein through to metabolites described by the central dogma might suggest that given protein abundances, RNA transcript levels are not informative about metabolite levels. However, there are many mechanisms which act contrary to the central dogma. For example, protein coding RNA transcripts are translated to proteins, and, conversely, proteins regulate transcription of DNA which synthesises RNA. Furthermore, proteins interact with each other and with metabolites in functional cellular biochemical processes.

There are several possible analyses which either exploit a pair of the data types (T+P, T+M, M+P) or all three simultaneously. The findings are not unanimous which is not surprising given different studies target different subsets of the transcriptome, proteome and metabolome which may or may not be directly related.

The highest correlation might be expected between the transcriptome and the proteome, though there is no consensus among scientists. Several studies investigate the correlation using a range of different samples ranging from yeast [55, 45] to human saliva [59]. Studies provide conflicting evidence: Greenbaum [52] suggest that up to 93% of 309 proteins found were at least present as mRNAs, and Gygi *et al.* [55] suggest protein abundances show 20-fold variability from mRNA levels. Greenbaum *et al.* [52] discuss reasons for this possible lack of correlation. One possible reason is rna interference. Another is the possibility that different profiling techniques target different expression regions of the gene. For example, probes for mRNAs are usually located in 3' UTR rather than protein coding regions. A recent study investigates this hypothesis by examining the correlation of the transcriptome and the proteome by correlating transcript expression at the exon level with peptide abundances [10].

As metabolites are highly sensitive to environmental exposures, including an individual's diet and microbiome, as well as the activity of other metabolites, they are not expected to be highly correlated with either transcript or protein abundance levels. However, integrated analysis of metabolites with transcript and/or proteins can still

provide insight into the biochemical processes and reactions which produce metabolites as their end products. Integrative strategies seek to map metabolite abundances as a function of gene transcripts or proteins, allowing the identification of genes mediating change in metabolic reactions. For example, see Cuperlovic-Culf *et al.* [29] for a review of integrated M+T and examples of applications of integrated M+T.

4.4 Integrated analysis of phenotype with at least two other sources of data

Many multifactorial complex diseases and phenotypes can be attributable to distinct or only partially overlapping risk factors in different individuals. At a genetic level this is most evident, such that even with large sample sizes, heterogeneity among disease cases (i.e. instances when different individuals have same disease as a result of different risk factors) can result in undetected multifactorial genetic risk factors. However, since complex diseases manifest through a group of diagnosable clinical phenotypes, it is likely that multifactorial risk factors collapse onto a smaller set of perturbed intermediate molecular pathways (namely transcript, protein and metabolic). This provides the motivation for integrating multiple data types and we present examples where these techniques have been used to assist identification of common sets of molecular pathways whose variation is associated with disease and aids the generation of causal phenotype/disease hypotheses. There are two ways in which data sources can be combined; either they can be analysed separately and these analyses compared or they can be analysed simultaneously and here we discuss both approaches.

4.4.1 Comparing genetic associations with different phenotypes

Genetic markers or regions significantly associated to disease or complex phenotypes can provide little (if any) functional context when considered in isolation. By exploiting the central dogma, eQTLs, pQTLs and mQTLs which overlap with another phenotype or indeed with each other can provide additional information which might help suggest putative mechanisms driving observed G-F associations. We use the term “overlapping” to mean that a region within a small window surrounding an eQTL, pQTL or mQTL also contains a locus or region associated with either another molecular trait or phenotype. Since mQTL and pQTL studies are only just beginning to be reported in literature here we largely discuss examples of eQTLs which overlap with a disease phenotype and their functional implications.

The results of the recent studies of human global gene expression discussed in section 4.2.3 have all been considered with respect to other phenotypes either with the same set of individuals [37, 51] or in an independent set of individuals [32, 113].

In a large GWA study Moffatt *et al.* [88], report a highly significant genetic association to asthma on chromosome 17q23 spanning nineteen genes including gene ORMDL3. The same markers are also reported as eQTLs significantly associated with all transcripts of the ORMDL3 gene in the study of gene expression by Dixon *et al.* [32]. Furthermore, after conditioning on the associated markers the expression trait for ORMDL3 is not significantly heritable, which suggests the expression of this gene is tightly regulated by genetic variation surrounding the eQTLs and provides a strong candidate gene with a causal role in at least one asthma pathway. In a similar way, Libioulle *et al.* [83] and Sladek *et al.* [118] combine their genetic association results with the global gene expression study of Dixon *et al.* [32] to discover a new candidate gene for Crohn’s disease and prioritise a candidate for Type-2 diabetes, respectively.

The eQTLs discovered by Schadt *et al.* [113] in their study of the genetic architecture of the human liver are compared with genetic associations with several common complex diseases including those investigated by the WTCCC [1]. The comparison with GWA hits for these diseases is natural since the liver is a metabolically active tissue and for many of them (particularly diabetes, obesity and atherosclerosis), the liver is thought to be instrumental in the manifestation of phenotype. Their eQTL study supports the prioritisation of one locus (RPS26) over another locus (ERBB3) both of which were newly identified as being associated with T1D in the WTCCC [1] study. In addition, they suggest candidates involved in the mechanisms causing CAD.

The study by Goring *et al.* [51] also investigates genetic associations a qualitative phenotype (high-density lipoprotein cholesterol (HDL-C)) which highly correlates with Cardiovascular disease. They find an overlapping eQTL (cis-acting on gene VNN1) with a region associated with HDL-C which supports the proposition of this gene

as a candidate for modifying cardiovascular disease risk via HDL-C. Emilsson *et al.* [37] do not report overlapping eQTLs with other phenotypes directly however they find that over 50% of expression traits in fat are highly correlated with obesity related traits (Body Mass Index and Percentage Body Fat). The genetic basis for this is supported with network and enrichment analysis.

There are few examples of overlapping protein or metabolite QTLs with clinical phenotype. This is perhaps because global metabolite and proteomic abundances are more sensitive to environmental factors and post-transcriptional and post-translational effects which makes detecting ‘pQTLs’ and ‘mQTLs’ more difficult. This might explain why there has been no replication of pQTL and mQTL findings and it makes comparisons across studies difficult. As the pQTL and mQTL studies become more widespread and reliable, we can expect comparisons between these regions and other regions which associate with both clinical (or other molecular) phenotypes to be of increasing importance in prioritising and suggesting causal phenotype hypotheses.

The examples which combine genetic, transcript and phenotypic data demonstrate the increase in power which can be gained by combining the analyses of multiple data types. Overlapping eQTLs and complex phenotype associated loci in humans have been successfully identified and used to rank of genetic candidates, add biological relevance to disease associated regions containing no genes or further localise large disease associated regions spanning multiple genes to a single candidate gene. However, this approach has only been used to suggest a single candidate gene for each of the diseases mentioned. This represents only a small fraction of the interacting pathways through which disease/complex phenotypes manifest. Furthermore, while such findings provide more biological context than simply a list of genetic associations and/or a gene list, considered in isolation they do not provide insight as to how variants regulate gene expression and what role these genes play in the determination of complex phenotypes. The discovery of overlapping pQTLs and mQTLs with eQTLs and disease associated regions might provide further insight although at best, can only suggest putative pathways.

A comprehensive understanding of the systems which give rise to overlapping associations necessitates the construction of focused study designs rather than the genome-wide discovery driven approaches (see section 5). For example by the use of exon arrays to detect alternatively spliced isoforms, tiling arrays to determine exact locations of transcription factor binding sites, and re-sequencing regions of association to identify causal genetic variants and their effects on gene regulation. These studies are costly, which further highlights the importance of improving specificity and sensitivity of analytical techniques used to suggest putative causal hypotheses from data collected at a genome-wide level.

4.4.2 Integrated Networks

Complex biological structures with interdependencies, temporal, spatial features can be conveniently (albeit crudely) represented using network models to interpret multiple sources of biological data. Although they are founded on huge assumptions, integrated networks in which edges are inferred between nodes both between and within specific different data types can suggest putative causal hypotheses consistent with the data and biological knowledge.

Network inference can be considered a model selection problem, but the entire space of possible networks is large (table 7) so fitting all models exhaustively is not computationally possible. Instead, the number of models tested is usually reduced by excluding those which are inconsistent with biological knowledge, or, local networks/pathways surrounding selected nodes might be tested rather than the entire network.

NETWORK TYPE	SIZE OF NETWORK SPACE WITH n NODES.
Undirected graphs	$\alpha_n = 2^{n(n-1)/2}$
Connected undirected graphs	$\beta_n = \alpha_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} \beta_k \alpha_{n-k}$
Directed Acyclic Graphs (DAGs)	$\gamma_n = \sum_{k=1}^n n(-1)^{k-1} \binom{n}{k} 2^{n(n-k)} \gamma_{n-k}$

Table 7: The dimensions of network space for different classes of network

These approaches are successfully applied to data from model organisms although there are few examples with human data. This is not surprising given that biological knowledge and data are more extensive in model organisms relative to human and that the underlying systems are less complex. The techniques used to infer integrated networks vary according to the data (and its dimension) and the study objectives. As molecular profiling of human samples (particularly of protein and metabolites) becomes more widespread, inference integrated networks will increasingly be used to aid interpretation and representation of the data.

Most techniques used to infer integrated networks use a mixture of the concepts (section 3); clearly they are founded on the concept of a network but they also draw on existing biological knowledge and the idea of a mapping from genotype to phenotype. Since there are no general frameworks commonly used, we describe some examples in table 8

DATA TYPES	KNOWLEDGE	NETWORK DESCRIPTION	REF.
G, T, F Obesity in Mice	Central Dogma used to restrict network space	Local causal pathways inferred using likelihood based methods to relate a single SNP, a gene expression trait and a complex trait (or another gene expression trait). Genetic loci and expression traits are selected for causal testing according to whether the expression trait and complex trait have overlapping QTLs	Schadt <i>et al.</i> [115]
G, T, P Gene Regulation in Yeast	Protein Interaction Data Bases.	Protein information used to construct a gene network apriori. Causal Pathways are sampled from this network given expression data and SNP data and target genes.	Tu <i>et al.</i> [130]
G, T, F Filamentous growth in Yeast	Central Dogma Protein Interaction Data Bases	The G, T and F used to derive a matrix of genetic influences (on T and F). This information is integrated with networks derived from data bases to make predictions about the effect of combinations of mutations on T and F	Carter <i>et al.</i> [17]
G, T, F Obesity, Diabetes and Atherosclerosis in Human			Chen <i>et al.</i> [20]
G, T, P TF-Binding (G-P) Gene regulation in Yeast	Protein Interaction Data Bases Central Dogma	Four different networks are constructed according different information used to infer them; a basic co-expression network (T,G) and three Bayesian Networks. The protein interaction information is incorporated via the use of Bayesian priors in the construction of the Bayesian Networks.	Zhu <i>et al.</i> [144]

Table 8: A selection of examples where integrated networks are inferred statistically. We briefly describe the data, knowledge and the network model used in each case.

Integrated analyses of phenotype with transcriptomic, proteomic and metabolomic data will play a larger role as these data types become more widespread and as the data improves in both resolution and precision. It is likely that both network and comparative approaches will be informative with the strategy of searching for overlapping eQTLs and phenotype associations being extended to compare all molecular and clinical phenotype associations. As yet there is no published study which maps global metabolite, protein and transcript abundances in the same set of individuals although this is a natural progression and extension of the techniques being currently developed.

4.5 Analysis of all data types across multiple species.

The study of phenotype in humans can be supplemented by similar studies in model organisms. At a genomic level, model organisms contribute knowledge to genome annotation via comparative genomics, but the use of model

organisms beyond their genomes is potentially much more powerful although methodologically less developed. Model organisms include a wide range of organisms, from those closely related to humans, such as primates, to those very distant such as yeast and bacteria. Incorporating several model organisms of varying degrees of relatedness can help distinguish between functional and spurious inferences.

Their use in research is widespread since many experiments cannot be conducted with humans due to ethical and/or practical reasons such as the longevity and inhomogeneous environmental backgrounds of humans.

The principle uses of model organisms are summarised in table 9.

Basic understanding of biological mechanisms

Life on earth has a set of common features and increased knowledge of one organism leads to increased understanding of related organisms. Many phenomena (for instance cell cycle and other fundamental cellular mechanisms) are much better studied in smaller, simpler organisms with a shorter generation time than in humans.

Homology modelling

This is closely related to understanding biology, but with a focus on particular level (e.g. proteins, networks, genes) and will be more amenable to exact representation and stochastic evolutionary models. The last years have seen this area flourish especially into the area of network inference and evolution (Sharan and Ideker, 2006). But clearly a series of areas, like motors, pattern formation, combined mathematical modelling in several species simultaneously etc. will follow.

Understanding mechanisms driving phenotypes

By focusing on phenotypes homologous, similar to or strongly correlated with the those observed in humans and relevant to a disease, it is hoped that an animal model for such a phenotype will have common underlying networks and related causes as the analogous human disease. A long series of phenotypes either are homologous or can be used as strongly correlated to the character of interest (Flint, nervous mice). For diseases, the ideal situation would be to have a homologous disease with a homologous cause. However, this is rarely if ever possible and a much weaker situation is used, where the model organisms develops a disease that has important similarities to the human disease but even these analyses rely on some degree of homology

Discovery of homologous or similar causes for a disease in model organisms

Most segregating nucleotides in human have an age less than a million years, hence the same disease mutation is very unlikely to occur in other species. The exception would be mutations subject to balancing selection, where the age of a mutant can be many million years and thus shared with closely related species. These cases are rare and the best that can be hoped for is a mutation in the same protein.

Transgenic experiments

The design of genetically modified model organisms, that are “locally human” with respect to a specific gene provides a powerful experimental tool. Human genetic material is placed in another organism, thus allowing experiments that would be impossible in humans. This has been crucial in many successes of finding and proving the molecular cause. However, given the large foreign genetic background and interconnectedness of genetic systems, there are also limits to such engineered models.

Table 9: The main uses of Model Organisms

Model organism studies of phenotype are in their early stages and as models of evolving networks and pathways emerge it is likely to contribute significantly to the understanding of human mechanisms driving complex phenotype. However, the use of closely related model organisms (including primates, dogs and cats) is severely criticised on ethical grounds. This has lead to a strong focus on finding computational or cell culture substitutes for the use of animal model organisms.

5 Functional Explanation

To gain a full functional understanding of the etiology of a complex phenotype involves (1) identifying the genetic, molecular, and environmental attributes that influence the phenotype, and (2) elucidating the biological pathway that fully defines the influence and describes how it occurs. While the analysis approaches that we have discussed above can be helpful in identifying features of (1) and (2), experimental validation is necessary for a comprehensive “functional explanation,” and we focus on attempts at such validation here. GWAS are the most high-profile means of a first stage attempt to accomplish the genetic component of (1); we therefore discuss key findings and the steps necessary to identify a causal genetic variant in section 5.1. Though this has proved extremely difficult, accomplishing (2) is a far more daunting task, involving a full understanding of systems biology and how all sources of biological and environmental variation interact (i.e. all-encompassing networks). We describe what this might entail in section 5.2.

5.1 Identifying Causal Genetic Variants

As mentioned, GWAS test for associations between genetic markers and a phenotype. To “confirm” a significant association between a marker and a phenotype from a GWAS, often the same marker (or a marker located physically very close) is typed in independent samples to see if the association is replicated. There are now over 300 such replicated disease associations [33]. However, determining the actual phenotype-influencing variant in a region with a replicated GWAS signal is not easy. Despite ongoing advances in the number of markers which can be typed for on a single array in GWAS, it is unlikely that even the most significant associations have a causative influence on any given phenotype. A signal of association at a genotyped marker is usually explained either by a single causal variant of lower frequency but with a larger effect located physically close to the marker or by multiple variants of differing frequencies (referred to as allelic heterogeneity) also located close to the marker. If (albeit unlikely) the causal variant and an associated genotyped marker are fully correlated (i.e. they have the same effect size and frequency), it is not possible to distinguish between them statistically and both markers would be followed up for functional characterisation. Resequencing can be employed to identify all forms of genetic variation in the region, including rare SNPs, copy number variation and rearrangements; alternatively, a new set of markers can be genotyped at a higher density. In practise often a combination of the two approaches is employed. Newly typed markers can either be tested for association in the standard way, or, the effect of existing variants can be adjusted for, such that the test establishes whether the newly typed variant accounts for phenotype variability independently of the other effects. Identifying a variant as being causal requires strong statistical evidence and functional characterisation. Without full functional characterisation, it cannot truly be considered causal.

Many disease studies are now reporting replicated finely mapped regions with strong evidence of association with disease (table 10) but very few can provide functional characterisation which explains the associations. In one such example [70], fine mapping strongly implicates a non-synonymous SNP Y402H as being causal for the development of age-related macular degeneration (AMD). This SNP induces a change in the complement factor H gene (CFH) product and although the CHF gene is known to be involved in complement activation and present in AMD lesions, more than two years after the initial report, the exact mechanism by which the variant protein increases disease risk remains to be unraveled. In another example [35], cell models have been used to follow up genetic associations and have pinpointed a functional role for risk variants. In this study, two out of eight correlated SNPs in the second intron of the FGFR2 gene that had been associated with increased risk of breast cancer were found to alter the binding affinity for transcription factors, giving rise to increased FGFR2 expression. Although still far from explaining how this variation increases breast cancer risk, these results serve as the first step in elucidating the functional implications of these genetic variants.

Although fine mapping is less widely reported for localising genetic variation associated with molecular traits (i.e. eQTLs, mQTLs and pQTLs), functional characterisation and identification of causal genetic variants can be supported by existing knowledge of the molecular trait. One example [51] is the localisation of the genetic variation which regulates the expression of gene VNN1. In the follow up study, the proximal promoter for this gene is resequenced in a subset of individuals from the original study and the six SNPs (out of 22) in this region which showed the most significant association with the VNN1 expression trait were then genotyped in all individuals in the study. The associations of five out of the six of these variants provided strong evidence of association. Determining which of these SNPs is the best causative candidate can be assessed using biological knowledge

catalogued. In this example, bioinformatic tools implicate one of the five SNPs as having functional implications on the ability of transcription factors to bind and is validated experimentally. How this SNP affects transcription factor binding remains to be answered. Another notable example is the genetic association of markers within the gene FADS1 with several metabolites [47]. In particular, a single SNP in a coding region of this gene explains upto 10% of the variation of the most strongly associated metabolite. They use information about known metabolic reactions to deduce that the minor allele at this variant causes a reduced efficiency of a specific fatty acid (Δ^5 saturase) reaction.

DISEASE	GENOMIC LOCUS	STRONGEST REPORTED	VARIATION	ALLELIC ODDS RATIO	REF.
Age-related macular degeneration	Chr1q31.3	SNP in an intron of the Complement Factor H (CHF) gene		4.6	[70]
Alzheimers disease	Chr19q13	SNP 14 kb distal to the APOE epsilon variant		40.1	[27]
Atrial fibrillation	Chr4q25	Two SNPs near the PITX2 gene, role in determining the left-right asymmetry of the heart		1.72 and 1.39	[53]
Breast cancer	Chr10q26.13	SNP in the second intron of the FGFR2 gene		1.26	[35]
Exfoliation glaucoma (XFG)	Chr15q24.1	Haplotypes formed by two non-synonomous SNPs in the LOXL1 gene (catalyzes the formation of elastin fibers, major component of lesions in XFG)		2.46 and 20.10 for the two SNPs	[128]
Cancer - multiple types	Chr8q24	At least 4 independent variants (3 SNPs and one haplotype) in a 0.7 Mb region upstream of the MYC gene found to associate with at least 4 different types of cancer		1.13-2.1	[46]
Obesity (BMI)	16q12.2	A SNP in an intron of the FTO gene		0.36kg/m ²	[41]
Schizophrenia and related psychoses	1q21	Rare 1.38 Mb deletion		14.83	[121]
Lung cancer and smoking behaviour	15q25	SNPs in and around nicotinic acetylcholine receptors (probably goes at least in part through smoking behaviour)		1.3 for lung cancer	[61, 127, 4]

Table 10: A selection of Genome-Wide Association study hits which are well replicated and provide good candidate genes being involved in manifestation of disease. In some cases there are hypotheses which propose possible mechanisms through which disease risk is increased.

Providing a functional characterisation of genetic variants associated with disease or a molecular trait remains an ongoing challenge which is perhaps not surprising given the complexity of the systems which lie on the path from genotype to phenotype. Comprehensive functional characterisation of a genetic variant will involve studies of epigenetic, genetic and environmental interactions together with full molecular dissection. This will inevitably lead to the rigorous testing of putative causal pathways.

5.2 Identifying Causal Pathways and Networks

Causal pathways and networks inferred statistically as we discussed in section 3.2.4 lack functional characterisation. Validating an entire global network is a huge task and usually specific pathways are prioritised and targeted for characterisation. Validation can be done via the perturbation of a system (e.g. genetic perturbation, gene silencing, gene knockouts) although even these studies do not provide a functional characterisation of what biological processes are driving the observed relationships; they merely provide support that the pathways are real. In

the network examples reported in table 8 most authors validate their findings via gene knockouts (in mice for the human example) but functional support to these findings is limited and founded on existing biological knowledge for example about a molecular structure, or a gene products role in well characterised biological processes. Like causal genetic variants, the identification of a causal pathway cannot be considered true until the mechanisms and molecular functions are fully characterised thereby fully annotating the links in a pathway with a biological process or reaction.

5.3 Forwards and Reverse Genetics

The majority of techniques we describe in this paper take a forwards genetics approach; data is gathered with reference to a phenotype and observations used to identify its causes. An alternative approach is reverse genetics where the starting point of a study is a set of genetic mutations and subsequent observations of phenotype are screened for differences. Both approaches aim to characterise the effect of genetic variation on phenotype.

Perturbation experiments in cell lines or model organisms are an example of reverse genetics, a genetic or molecular adjustment is made (e.g. a gene is knocked out) and the consequences on phenotype are observed. They are useful for validation and refinement of hypotheses which are important to direct the focus of functional studies but are not likely to be informative at a fine level for human systems. This is because human mechanisms are often disrupted by very subtle effects rather than severe perturbations of the scale induced by gene knockouts, furthermore the functional effects of a genetic variant may take a long time to manifest.

A forwards genetic approach to elucidate disease mechanisms in humans directly might be to select individuals who carry a specific genetic variant and subsequently measure a range of different phenotypes (molecular, morphological, anthropometric, behavioural) over a time period. Although mutations cannot be engineered in humans, genetic variation is present naturally in a population and could be exploited systematically.

6 Conclusion

In this review, we distinguish data, concepts, and analyses as the major components of studies of biological variation. Ideally they should be clearly outlined prior to performing an experiment or gathering data, although in practise many researchers observe data before giving consideration to the concepts and analytical techniques to interpret it. In addition, studies can be complemented by a variety of other data types or information which can be interpreted in an intuitive informal way rather than according to the framework we describe. However, it is our hope that this framework still describes major features of present research and as such, is useful in extracting some unity in what can seem bewilderingly complex.

Full understanding and predictive modelling of biological systems are the ultimate goals of research in the biosciences but the global genome-wide studies describe systems of a size that cannot be modelled to this level in the foreseeable future. While they have been, and will continue to be, tremendously useful in coarse functional assignments of genes, they cannot lead to detailed understanding of biological systems. This level of understanding will inevitably be the result of bottom-up approaches which provide a more detailed understanding of smaller systems or fewer genes.

The global integrative strategies we discussed in this review use statistical models to interpret observed high-throughput biological data. The ability to make causal statements on this basis is limited by biological and technical noise inherent in the data types and any findings require careful validation and subsequent functional explanation. Systems biology approaches could be used to provide this explanation; each time a candidate causal pathway is identified statistically from data, systems biology would aim to provide a physical explanation by constructing a physical model which can then also be used for prediction. This is a very ambitious goal and it could be many years before this achieved. Currently systems biology is being done on a small scale, whether it can be done for large systems and how difficult it will be is impossible to predict.

Directly or indirectly, most funding for research in the biosciences is motivated by the hope of alleviating human disease. The overall trajectory of disease research might be:

Association Mapping → Fine Mapping → Functional Interpretation → Intervention/ Drug Development (9)

The first two steps are routinely employed within the framework we describe. Functional interpretation is attempted by integrative studies and systems biology but both of these techniques are still too high level to provide full functional explanations at a molecular or atomic level. This level of explanation is achieved largely by highly focused laboratory experiments with statistics playing a small role relative to the techniques employed to analyse global, high throughput biological data sets. The final step 'Intervention/ Drug Development' is already a well established field despite the fact that there are very few functional explanations of complex disease which start with genetics. Consequently, it requires huge resources.

Functional interpretation is possibly the hardest stage in 9, and it is likely to be a slow process since progress requires many experiments focused on a few relevant components, complemented with traditional biochemical analysis. Representation of these systems requires models, but they will have to describe how small systems can be interpreted in detail. Whilst there is still much work to be done in association mapping and integrative genomics to prioritise candidates for functional characterisation, ultimately the long series of large scale genome wide studies will have to come to a dead end since it cannot provide full functional explanations.

The data types upon which integrative genomics and other studies of global high throughput data are founded, are quantities measured in the cell. With the exception of phenotype, information and data involving higher levels play a small, if any role. Given the fundamental impossibility of defining a function which maps genotype to phenotype function presently and in the foreseeable future, all observations which can help to characterise this function are of value. Many new data types, such as medical imaging, relating to tissues and organs will be of increasing relevance and already have proved valuable in the study of model organisms.

The necessity of large integrated approaches is changing the biosciences and how the individual researcher conducts research. Two decades ago computers were irrelevant for most researchers in the biosciences. A decade ago molecular evolution was a specialised area, now it is of general use in genome studies. Just over 5 years ago the same was true for population studies, now association mapping is ubiquitous. We are presently seeing the rise of high throughput studies. The near future will probably see mathematical modelling being important to everyone.

References

- [1] (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661–678.
- [2] Abe, N. and Mamitsuka, H. (1998). Predicting protein secondary structure using stochastic tree grammars. *Machine Learning*, **29**, 275–301.
- [3] Aceto, L. (2003). Some of my favorite results in classic process algebra. *Bulletin of the EATCS*, **81**, 89–108.
- [4] Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y.-Y., Chen, W. V., Shete, S., Spitz, M. R., and Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*, **40**(5), 616–622.
- [5] Baeten, J. (2006). A brief history of process algebra. *Theoretical Computer Science*, **335**(2-3), 131–146.
- [6] Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**(10), 781–791.
- [7] Bard, J. and Rhee, S. (2004). Ontologies in biology: design, applications, and future challenges. *Nat Rev Genetics*, **5**, 213–222.
- [8] Bilder, R. (2008). Phenomics: building scaffolds for biological hypotheses in the post-genomic era. *Biol Psychiatry*, **63**(5), 439–440.
- [9] Bird, A. (2007). Perceptions of epigenetics. *Nature*, **447**(7143), 396–398.
- [10] Bitton, D. A., Okoniewski, M. J., Connolly, Y., and Miller, C. J. (2008). Exon level integration of proteomics and microarray data. *BMC Bioinformatics*, **9**, 118+.
- [11] Bodenreicher, O. and Stevens, R. (2004). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, **7**(3), 256–274.
- [12] Bogue, M. A., Grubb, S. C., Maddatu, T. P., and Bult, C. J. (2007). Mouse Phenome Database (MPD). *Nucleic acids research*, **35**(Database issue).
- [13] Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., and Thorsson, V. (2006). The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, **7**(5).
- [14] Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.
- [15] Calder, M., Gilmore, S., and Hillston, J. (2005). Automatically deriving ODEs from process algebra models of signalling pathways. *Proceedings of Comp Methods in Systems Bio*, pages 204–215.
- [16] Calder, M., Gilmore, S., and Hillston, J. (2006). Modelling the influence of RKP on the ERK signalling pathway using the stochastic process algebra PEPA. *Transactions on Computational Systems Biology VII*, Springer, **4230**, 1–13.
- [17] Carter, G. W., Prinz, S., Neou, C., Shelby, P. J., Marzolf, B., Thorsson, V., and Galitski, T. (2007). Prediction of phenotype and gene expression for combinations of mutations. *Mol Syst Biol*, **3**.
- [18] Chaerkady, R. and Pandey, A. (2008). Applications of proteomics to lab diagnosis. *Annual review of pathology*, **3**, 485–498.
- [19] Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., and O’Connell, P. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet*, **362**(9381), 362–369.
- [20] Chen, Y., Zhu, J., Lum, P. Y., Yang, X., Pinto, S., Macneil, D. J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S. K., Leonardson, A., Castellini, L. W., Wang, S., Champy, M.-F., Zhang, B., Emilsson, V., Doss, S., Ghazalpour, A., Horvath, S., Drake, T. A., Lusk, A. J., and Schadt, E. E. (2008). Variations in DNA elucidate molecular networks that cause disease. *Nature*.

- [21] Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K. Y., Morley, M., and Spielman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, **33**(3), 422–425.
- [22] Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M., and Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**(7063), 1365–1369.
- [23] Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
- [24] Churchill, G. (1989). Stochastic models for heterogenous DNA sequences. *Bulletin of Mathematical Biology*, **51**(1), 79–94.
- [25] Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, pages 0709013104+.
- [26] Clark, T. A., Sugnet, C. W., and Ares, M. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**(5569), 907–910.
- [27] Coon, K. D., Myers, A. J., Craig, D. W., Webster, J. A., Pearson, J. V., Lince, D. H., Zismann, V. L., Beach, T. G., Leung, D., Bryden, L., Halperin, R. F., Marlowe, L., Kaleem, M., Walker, D. G., Ravid, R., Heward, C. B., Rogers, J., Papassotiropoulos, A., Reiman, E. M., Hardy, J., and Stephan, D. A. (2007). A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer’s disease. *The Journal of clinical psychiatry*, **68**(4), 613–618.
- [28] Cornishbowden, A., Cardenas, M., Letelier, J., Sotoandrade, J., and Abarzua, F. (2004). Understanding the parts in terms of the whole. *Biology of the Cell*, **96**(9), 713–717.
- [29] Cuperlovic-Culf, M., Belacel, N., and Culf, A. (2008). Integrated analysis of transcriptomics and metabolomics profiles. *Expert Opinion on Medical Diagnostics*, pages 497–509.
- [30] Dawn and Barrett, J. H. (2005). Genetic linkage studies. *The Lancet*, **366**(9490), 1036–1044.
- [31] de Koning, D. J. and Haley, C. S. (2005). Genetical genomics in humans and model organisms. *Trends Genet*, **21**(7), 377–381.
- [32] Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, M. G., Abecasis, G. R., and Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, **39**(10), 1202–1207.
- [33] Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, **456**(7223), 728–731.
- [34] Dunn, W. B. and Ellis, D. (2005). Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, **24**(4), 285–294.
- [35] Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struwing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., Meyer, K. B., Haiman, C. A., Kolonel, L. K., Henderson, B. E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C.-Y., Wu, P.-E., Wang, H.-C., Eccles, D., Evans, G. D., Peto, J., Fletcher, O., Johnson, N., Seal, S., Stratton, M. R., Rahman, N., Chenevix-Trench, G., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K.-Y., Noh, D.-Y., Ahn, S.-H., Hunter, D. J., Hankinson, S. E., Cox, D. G., Hall, P., Wedren, S., Liu, J., Low, Y.-L., Bogdanova, N., Schürmann, P., Dörk, T., Tollenaar, R. A. E. M., Jacobi, C. E., Devilee, P., Klijn, J. G. M., Sigurdson, A. J., Doody, M. M., Alexander, B. H., Zhang, J., Cox, A., Brock, I. W., Macpherson, G., Reed, M. W. R., Couch, F. J., Goode, E. L., Olson, J. E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J. L., McCreddie, M., Southey, M., Giles, G. G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko, Y.-D., Spurdle, A. B., Beesley, J., Chen, X., Mannermaa, A., Kosma, V.-M., Kataja, V., Hartikainen, J., Day, N. E., Cox, D. R., and Ponder, B. A. J. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*.

- [36] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**(25), 14863–14868.
- [37] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, B. G., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G. H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K. P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H. G., Stefansson, T., Leifsson, B. G., Thorsteinsdottir, U., Lamb, J. R., Gulcher, J. R., Reitman, M. L., Kong, A., Schadt, E. E., and Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature*.
- [38] Ernst, J., Vainas, O., Harbison, C. T., Simon, I., and Bar-Joseph, Z. (2007). Reconstructing dynamic regulatory maps. *Mol Syst Biol*, **3**.
- [39] Fan, X., Bai, J., and Shen, P. (2005). Diagnosis of breast cancer using hplc metabonomics fingerprints coupled with computational methods. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 6081–6084.
- [40] Felsenstein, J. and Churchill, G. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, **13**, 93–104.
- [41] Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M.-R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N., Doney, A. S., Morris, A. D., Smith, G. D., The, Hattersley, A. T., and McCarthy, M. I. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*, **316**(5826), 889–894.
- [42] Freimer, N. and Sabatti, C. (2003). The Human Phenome Project. *Nat Genet*, **34**(1), 15–21.
- [43] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, **303**(5659), 799–805.
- [44] Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U., and Shaprio, E. (2005). Genomic variability within an organism exposes its cell lineage tree. *PLoS Compu Bio*, **1**(5), e50.
- [45] Futcher, B., Latter, G. I., Monardo, P., Mclaughlin, C. S., and Garrels, J. I. (1999). A sampling of the yeast proteome. *Mol. Cell. Biol.*, **19**(11), 7357–7368.
- [46] Ghossaini, M., Song, H., Koessler, T., Al, Kote-Jarai, Z., Driver, K. E., Pooley, K. A., Ramus, S. J., Kjaer, S. K., Hogdall, E., Dicioccio, R. A., Whittemore, A. S., Gayther, S. A., Giles, G. G., Guy, M., Edwards, S. M., Morrison, J., Donovan, J. L., Hamdy, F. C., Dearnaley, D. P., Arden-Jones, A. T., Hall, A. L., O’Brien, L. T., Gehr-Swain, B. N., Wilkinson, R. A., Brown, P. M., Hopper, J. L., Neal, D. E., Pharoah, P. D., Ponder, B. A., Eeles, R. A., Easton, D. F., Dunning, A. M., and For (2008). Multiple Loci With Different Cancer Specificities Within the 8q24 Gene Desert. *J. Natl. Cancer Inst.*, **100**(13), 962–966.
- [47] Gieger, C., Geistlinger, L., Altmaier, E., Hrabé, Kronenberg, F., Meitinger, T., Mewes, H.-W., Wichmann, Weinberger, K. M., Adamski, J., Illig, T., and Suhre, K. (2008). Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, **4**(11), e1000282+.
- [48] Gkoutos, O., Green, E., Mallon, A., Hancock, J., and Davidson, D. (2004). Building mouse phenotype ontologies. *Pacific Symposium on Biocomputing*, **9**, 178–189.
- [49] Goldman, N., Thorne, J., and Jones, D. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.
- [50] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**(5439), 531–537.
- [51] Goring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., Jowett, J. B., Abraham, L. J., Rainwater, D. L., Comuzzie, A. G., Mahaney, M. C., Almasy, L., Maccluer, J. W., Kissebah,

- A. H., Collier, G. R., Moses, E. K., and Blangero, J. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet*, **39**(10), 1208–1216.
- [52] Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*, **4**(9).
- [53] Gudbjartsson, D. F., Arnar, D. O., Helgadóttir, A., Gretarsdóttir, S., Holm, H., Sigurdsson, A., Jonasdóttir, A., Baker, A., Thorleifsson, G., Kristjansson, K., Palsson, A., Blondal, T., Sulem, P., Backman, V. M., Hardarson, G. A., Palsdóttir, E., Helgason, A., Sigurjonsdóttir, R., Sverrisson, J. T., Kostulas, K., Ng, M. C. Y., Baum, L., So, W. Y., Wong, K. S., Chan, J. C. N., Furie, K. L., Greenberg, S. M., Sale, M., Kelly, P., Macrae, C. A., Smith, E. E., Rosand, J., Hillert, J., Ma, R. C. W., Ellinor, P. T., Thorgeirsson, G., Gulcher, J. R., Kong, A., Thorsteinsdóttir, U., and Stefansson, K. (2007). Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*.
- [54] Gudmundsson, J., Sulem, P., Steinthorsdóttir, V., Bergthorsson, J., Thorleifsson, G., Manolescu, A., Rafnar, T., Gubjartsson, D., Agnarsson, B., Baker, A., Sigurdsson, A., Benediksdóttir, K., Jakobsdóttir, M., Blondal, T., Stacey, S., Helgason, A., Gunnarsdóttir, S., Olafsdóttir, A., Kristinsson, K., Birgisdóttir, B., Kostic, J., Gosh, S., and et al., S. T. (2007). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet*, **39**, 977–983.
- [55] Gygi, S. P., Rochon, Y., Franza, R. B., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.*, **19**(3), 1720–1730.
- [56] Hall, D. A., Ptacek, J., and Snyder, M. (2007). Protein microarray technology. *Mechanisms of ageing and development*, **128**(1), 161–167.
- [57] Hanash, S. M., Baier, L. J., McCurry, L., and Schwartz, S. A. (1986). Lineage-related polypeptide markers in acute lymphoblastic leukemia detected by two-dimensional gel electrophoresis. *Proceedings of the National Academy of Sciences of the United States of America*, **83**(3), 807–811.
- [58] Helgason, A., Hrafnkelsson, B., Gulcher, J., Ward, R., and Stefansson, K. (2003). A populationwide coalescent analysis of icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet*, **72**(5), 1370–1388.
- [59] Hu, S., Li, Y., Wang, J., Xie, Y., Tjon, K., Wolinsky, L., Loo, R. R., Loo, J. A., and Wong, D. T. (2006). Human saliva proteome and transcriptome. *J Dent Res*, **85**(12), 1129–1133.
- [60] Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23**, 183–201.
- [61] Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., Chen, C., Goodman, G., Field, J. K., Liloglou, T., Xinarianos, G., Cassidy, A., Mclaughlin, J., Liu, G., Narod, S., Krokan, H. E., Skorpén, F., Elvestad, M. B., Hveem, K., Vatten, L., Linseisen, J., Clavel-Chapelon, F., Vineis, P., Bueno-De-Mesquita, B. H., Lund, E., Martínez, C., Bingham, S., Rasmuson, T., Hainaut, P., Riboli, E., Ahrens, W., Benhamou, S., Lagiou, P., Trichopoulos, D., Holcatova, I., Merletti, F., Kjaerheim, K., Agudo, A., Macfarlane, G., Talamini, R., Simonato, L., Lowry, R., Conway, D. I., Znaor, A., Healy, C., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Matsuda, F., Blanche, H., Gut, I., Heath, S., Lathrop, M., and Brennan, P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, **452**(7187), 633–637.
- [62] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- [63] International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
- [64] Jacob, F., Perrin, D., Sanchez, C., and Monod, J. (1960). [operon: a group of genes with the expression coordinated by an operator.]. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, **250**, 1727–1729.
- [65] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, R. J., Vanliere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez,

- J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., and Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**(7181), 998–1003.
- [66] Jansen, R. C. and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics*, **17**(7), 388–391.
- [67] Kaddurah-Daouk, R., Mcevoy, J., Baillie, R. A., Lee, D., Yao, J. K., Doraiswamy, P. M., and Krishnan, K. R. R. (2007). Metabolomic mapping of atypical antipsychotic effects in schizophrenia. *Molecular Psychiatry*, **aop**(current).
- [68] Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science (New York, N.Y.)*, **245**(4922), 1073–1080.
- [69] Kingman, J. (1982). The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.
- [70] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., Sangiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**(5720), 385–389.
- [71] Klose, J. and Kobalz, U. (1995). Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis*, **16**(6), 1034–1059.
- [72] Knudsen, B. and Hein, J. (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.
- [73] Kou, S., Sunney, X., and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 469–506.
- [74] Krogh, A., Brown, M., Saira Mian, I., Sjolander, K., and Haussler, D. (1994). Hidden markov models in computational biology: applications to protein modeling. *J Mol Biol*, **235**, 1501–1531.
- [75] Kulzer, F. and Orrit, M. (2004). Single-molecule optics. *Annual Review of Physical Chemistry*, **55**(1), 585–611.
- [76] Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T. J. J., Sladek, R., and Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. *Nat Genet*.
- [77] Kwiatkowska, M., Heath, J., and Gaffney, E. (2006). Simulation and verification for computational modelling of signalling pathways. *Proc Winter Simulation Conference*.
- [78] Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A*, **84**(8), 2363–2367.
- [79] Lange, E., Robbins, C., Gillanders, E., Zheng, S., Xu, J., Wang, Y., White, K., Chang, B., Ho, L., Trent, J., Carpten, J., Isaacs, W., and Cooney, K. (2007). Fine-mapping the putative chromosome 17q21-22 prostate cancer susceptibility gene to a 10 cM region based on linkage analysis. *Hum Genet*, **121**, 49–55.
- [80] Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- [81] Lawrence, N., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. *Advances in Neural Information and Processing Systems (NIPS 20)*.
- [82] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., Macdonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., Mcintosh, T. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y.-H., Frazier, M. E., Scherer, S. W., Strausberg, R. L., and Venter, C. J. (2007). The diploid genome sequence of an individual human. *PLoS Biology*, **5**(10), e254+.
- [83] Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A., Demarche, B., Gut, I., Heath, S., Foglio, M., Liang, L., Laukens, D., Mni, M., Zelenika, D., Van Gossum, A., Rutgeerts, P., Belaiche, J., Lathrop, M., and Georges, M. (2007). Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS genetics*, **3**(4).

- [84] MAQC (????).
- [85] Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*.
- [86] Mateos-Cáceres, P. J., García-Méndez, A., López Farré, A., Macaya, C., Núñez, A., Gómez, J., Alonso-Orgaz, S., Carrasco, C., Burgos, M. E., de Andrés, R., Granizo, J. J., Farré, J., and Rico, L. A. (2004). Proteomic analysis of plasma from patients during an acute coronary syndrome. *Journal of the American College of Cardiology*, **44**(8), 1578–1583.
- [87] Melzer, D., Perry, J. R., Hernandez, D., Corsi, A. M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., Rafiq, S., Simon-Sanchez, J., Lango, H., Scholz, S., Weedon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., Britton, A., Henley, W., van de Leemput, J., Li, R., Newman, A. B., Tranah, G., Harris, T., Panicker, V., Dayan, C., Bennett, A., McCarthy, M. I., Ruukonen, A., Jarvelin, M. R., Guralnik, J., Bandinelli, S., Frayling, T. M., Singleton, A., and Ferrucci, L. (2008). A genome-wide association study identifies protein quantitative trait loci (pqtls). *PLoS genetics*, **4**(5).
- [88] Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., Heinzmann, A., Simma, B., Frischer, T., Willis-Owen, S. A. G., Wong, K. C. C., Illig, T., Vogelberg, C., Weiland, S. K., von Mutius, E., Abecasis, G. R., Farrall, M., Gut, I. G., Lathrop, M. G., and Cookson, W. O. C. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*.
- [89] Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J. W., Sachs, A., and Schadt, E. E. (2004). Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*, **75**(6), 1094–1105.
- [90] Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**(7001), 743–747.
- [91] Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**(5746), 321–324.
- [92] Nielsen, R., Bustamante, C., Clark, A., Gnanowski, S., Sackton, T., Hubisz, M., Fledel-Alon, A., Tanenebaum, D., Civello, D., White, T., Sninsky, J., Adams, M., and Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, **3**(6), e170.
- [93] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*.
- [94] Olson, J. M., Witte, J. S., and Elston, R. C. (1999). Genetic mapping of complex traits. *Statistics in medicine*, **18**(21), 2961–2981.
- [95] Pastinen, T., Ge, B., and Hudson, T. J. (2006). Influence of human genome polymorphism on gene expression. *Hum Mol Genet*, **15 Spec No 1**.
- [96] Pedersen, J. and Hein, J. (2003). Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, **19**(2), 219–227.
- [97] Phillips, A. and Cardelli, L. (2004). A correct abstract machine for the stochastic pi-calculus. *BioConcur: Electronic Notes in Theoretical Computer Science*.
- [98] Phillips, A., Cardelli, L., and Castagna, G. (2006). A graphical representation for biological processes in the stochastic pi-Calculus. *Transactions on Computational Systems Biology VII, Springer*, **4230**, 123–152.
- [99] Ponting, C. P. P. (2008). The functional repertoires of metazoan genomes. *Nature reviews. Genetics*.
- [100] Posch, M. G., Berger, F., Perrot, A., and Ozcelik, C. (2008). We need a detailed phenome in the phenomenon of genetics and congenital heart disease. *Journal of medical genetics*, **45**(5).
- [101] Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 504–509.
- [102] Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, **69**, 1–14.

- [103] Puzstai, L. (2008). Current Status of Prognostic Profiling in Breast Cancer. *Oncologist*, **13**(4), 350–360.
- [104] Rakha, A. E., El-Sayed, E. M., Reis-Filho, S. J., Ellis, and O, I. (2008). Expression profiling technology: its contribution to our understanding of breast cancer. *Histopathology*, **52**(1), 67–81.
- [105] Raychaudhuri, S. (2006). *Computational text analysis for functional genomics and bioinformatics*. Oxford University Press.
- [106] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, D. T., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., Macdonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, **444**(7118), 444–454.
- [107] Rozen, S., Cudkowicz, M. E., Bogdanov, M., Matson, W. R., Kristal, B. S., Beecher, C., Harrison, S., Vouros, P., Flarakos, J., Vigneau-Callahan, K., Matson, T. D., Newhall, K. M., Beal, F. M., Brown, R. H., and Kaddurah-Daouk, R. (2005). Metabolomic analysis and signatures in motor neuron disease. *Metabolomics*, **1**(2), 101–108.
- [108] Sabatine, M. S., Liu, E., Morrow, D. A., Heller, E., Mccarroll, R., Wiegand, R., Berriz, G. F., Roth, F. P., and Gerszten, R. E. (2005). Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, **112**(25), 3868–3875.
- [109] Sakakibara, Y., Brown, M., Hughey, R., Saira Mian, I., Sjolander, K., Underwood, R., and Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res*, **22**(23), 5112–5120.
- [110] Sandelin, A. and Wasserman, W. (2004). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, **338**(2), 107–215.
- [111] Sanguinetti, Guido, Lawrence, Neil, D., Rattray, and Magnus (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**(22), 2775–2781.
- [112] Savageau, M. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Addison Wesley Publishing Company.
- [113] Schadt, E., Molony, C., and Ulrich, R. (2008). Mapping the Genetic Architecture of Gene Expression in the Human Liver. *PLoS Biology*, **6**, 1020–1032.
- [114] Schadt, E. E., Monks, S. A., Drake, T. A., Lusk, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**(6929), 297–302.
- [115] Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusk, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, **37**(7), 710–717.
- [116] Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754–764.
- [117] Segal, E., Shapira, M., Regev, A., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: Discovering regulatory modules and their condition specific regulators from gene expression data. *Nature Genetics*, **34**.
- [118] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*.
- [119] Soreq, L., Gilboa-Geffen, A., Berrih-Aknin, S., Lacoste, P., Darvasi, A., Soreq, E., Bergman, H., and Soreq, H. (????). [identifying alternative hyper-splicing signatures in mg-thymoma by exon arrays. *PLoS ONE*, (6).

- [120] Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Li, P., Eising, P. E., Brown, P. O., Børresen-Dale, A. L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, **100**(14), 8418–8423.
- [121] Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J. E., Hansen, T., Jakobsen, K. D., Muglia, P., Francks, C., Matthews, P. M., Gylfason, A., Halldorsson, B. V., Gudbjartsson, D., Thorgeirsson, T. E., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Bjornsson, A., Mattiasdottir, S., Blondal, T., Haraldsson, M., Magnusdottir, B. B., Giegling, I., Möller, H.-J. a., Hartmann, A., Shianna, K. V., Ge, D., Need, A. C., Crombie, C., Fraser, G., Walker, N., Lonnqvist, J., Suvisaari, J., Tuulio-Henriksson, A., Paunio, T., Touloupoulou, T., Bramon, E., Di Forti, M., Murray, R., Ruggeri, M., Vassos, E., Tosato, S., Walshe, M., Li, T., Vasilescu, C., Mäkilä, P., Helle, T. W., Wang, A. G., Ullum, H., Djurovic, S., Melle, I., Olesen, J., Kiemene, L. A., Franke, B., Sabatti, C., Freimer, N. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., Andreassen, O. A., Ophoff, R. A., Georgi, A., Rietschel, M., Werge, T., Petursson, H., Goldstein, D. B., Nöthen, M. M., Peltonen, L., Collier, D. A., St Clair, D., and Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*.
- [122] Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S. E., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genet*, **1**(6).
- [123] Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**(5813), 848–853.
- [124] Stumpf, M., Thorne, T., de Silva, E., Stewart, R., An, H., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *PNAS*, **105**, 6959–6964.
- [125] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat Genet*, **22**(3), 281–285.
- [126] The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
- [127] Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsater, A., Flex, A., Aben, K. K., de Vegt, F., Mulders, P. F., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., van Rij, A. M., Dieplinger, B., Haltmayer, M., Mayordomo, J. I., Kiemene, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfinsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, **452**(7187), 638–642.
- [128] Thorleifsson, G., Magnusson, K. P., Sulem, P., Walters, G. B., Gudbjartsson, D. F., Stefansson, H., Jonsson, T., Jonasdottir, A., Jonasdottir, A., Stefansdottir, G., Masson, G., Hardarson, G. A., Petursson, H., Arnarsson, A., Motallebipour, M., Wallerman, O., Wadelius, C., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., Jonsson, F., and Stefansson, K. (2007). Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science (New York, N.Y.)*, **317**(5843), 1397–1400.
- [129] Thorsen, K., Sorensen, K. D., Brems-Eskildsen, A. S., Modin, C., Gaustadnes, M., Hein, A.-M. K., Kruhoffer, M., Laurberg, S., Borre, M., Wang, K., Brunak, S., Krainer, A. R., Topping, N., Dyrskjot, L., Andersen, C. L., and Orntoft, T. F. (2008). Alternative splicing in colon, bladder, and prostate cancer identified by exon-array analysis. *Mol Cell Proteomics*, pages M700590–MCP200+.
- [130] Tu, Z., Wang, L., Arbeitman, M. N., Chen, T., and Sun, F. (2006). An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, **22**(14).

- [131] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerckhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.
- [132] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooshef, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.
- [133] Voight, B., Kudaravalli, S., Wen, X., and Pritchard, J. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, **4**(3), e72.
- [134] Wang, G.-S. and Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, **8**(10), 749–761.
- [135] Wasserstrom, A., Adar, R., Shefer, G., Frumkin, D., Itzkovitz, S., Stern, T., Shur, I., Zangi, L., Kaplan, S., Harmelin, A., Reisner, Y., Benayahu, D., Tzahor, E., Segal, E., and Shapiro, E. (2008). Reconstruction of cell lineage trees in mice. *PLoS ONE*, **3**(4), e1939.
- [136] Weckwerth, W. (2003). Metabolomics in systems biology. *Annu Rev Plant Biol*, **54**, 669–689.
- [137] Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A., and Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences*, **98**(3), 1176–1181.
- [138] Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., Mcguire, A., He, W., Chen, Y.-J., Makhi-jani, V., Roth, T. G., Gomes, X., Tartaro, K., Niaz, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M.,

- Gibbs, R. A., and Rothberg, J. M. (2008). The complete genome of an individual by massively parallel dna sequencing. *Nature*, **452**(7189), 872–876.
- [139] Wilkinson, D. (2006). *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC.
- [140] Xiang, Y. and Kato, T. (2006). Use of proteomics in analysis of autoimmune diseases. *Lupus*, **15**(7), 431–435.
- [141] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, **4**(1).
- [142] Zhang, J., Goodlett, D. R., Quinn, J. F., Peskind, E., Kaye, J. A., Zhou, Y., Pan, C., Yi, E., Eng, J., Wang, Q., Aebersold, R. H., and Montine, T. J. (2005). Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease. *Journal of Alzheimer's disease : JAD*, **7**(2).
- [143] Zhu, J., Lum, P. Y., Lamb, J., GuhaThakurta, D., Edwards, S. W., Thieringer, R., Berger, J. P., Wu, M. S., Thompson, J., Sachs, A. B., and Schadt, E. E. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*, **105**(2-4), 363–374.
- [144] Zhu, J., Zhang, B., Smith, E. N. N., Drees, B., Brem, R. B. B., Kruglyak, L., Bumgarner, R. E. E., and Schadt, E. E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*.