

Studies of Global Biological and their Contribution to Complex Phenotype *or* From Integrative Genomics towards Functional Characterisation.

November 29, 2008

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. Traditionally, studies were 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	Transcriptomic Data	5
2.3	Proteomic Data	7
2.4	Metabolomic Data	8
2.5	Epigenomic Data	9
2.6	Phenomic Data	9
2.7	Other Data Sources	10
3	Concepts	13
3.1	A Mapping from Genome G to Phenome F	13
3.2	Networks	16
3.2.1	Biological Networks	16
3.2.2	Physical Networks	17
3.2.3	Statistical Networks	17
3.2.4	Biological Interpretation of Statistical Networks	19
3.3	Genealogical Relationships	20
3.3.1	Genealogies relating cells in an individual	20
3.3.2	Genealogies relating individuals of a population	21
3.3.3	Genealogies relating species	21
3.4	Knowledge	22
3.5	Hidden Structures	23
4	Analyses	24
4.1	Analysis of single sources of data	24
4.1.1	Genomic Variation Data	24
4.1.2	Human Genetic Variation Data (G)	25
4.1.3	Molecular Phenotypes	26
4.2	Analysis of phenotype with another source of data	27
4.2.1	Analysis with clinical phenotype with genetic data ($G+F$)	27
4.2.2	Analysis with clinical phenotype and molecular phenotype ($F + T$), ($F + P$), ($F + M$)	29
4.2.3	Analysis with molecular phenotype with genetic data ($G + T, P$ or M)	31
4.3	Analysis with two molecular phenotypes ($T+P, T+M, M+P$)	32
4.3.1	Correlation between the Transcriptome and Proteome ($T+P$)	32
4.3.2	Correlation between the Transcriptome or Proteome and Metabolome ($T+M, P+M$)	32
4.4	Integrated analysis of complex phenotype with at least two other sources of data	33
4.4.1	Comparing genetic associations with different phenotypes. ($G + F$ with $G + T, G + P, G + M$)	33
4.4.2	Integrated Networks	34
4.5	Analysis of all types of biological data across multiple species	35
5	Focused Studies	36
6	Conclusion	36

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.

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 at least one of 5 concepts (C1-C5). They are accepted and used so frequently that they are often taken for granted and used without question. Firstly, the assumption that a mapping from genotype to phenotype exists 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 is genealogies. They can be considered at three levels 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.

C1 A Mapping from Genotype to Phenotype

C2 Networks

C3 Genealogical Relationships

C4 Biological Knowledge

C5 Hidden Structures

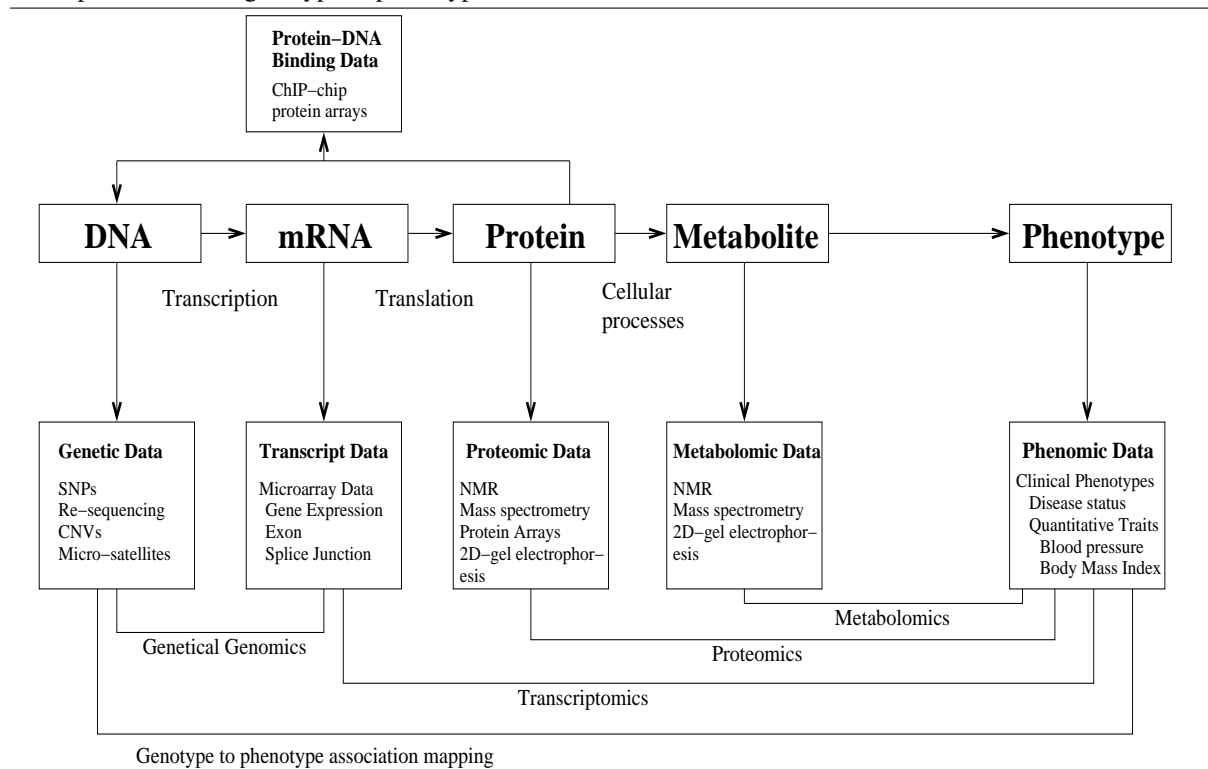
Fourth is the concept of knowledge, it is difficult to define and represent but crucial since all studies are founded on it to some degree. At a foundational level, knowledge shapes study design and analytical strategies. At a more specific level, knowledge can be informative for either the construction 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. JOTUN MODIFY THIS SENTENCE ACCORDING TO WHAT YOU HAVE WRITTEN IN THE TEXT?

Data and concepts are combined to construct models. These can be used to interpret and analyse data, and different combinations of data 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 and 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 and this 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 which also play a vital role in many of the processes from genotype to phenotype. See section 2.5



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 (Ponting, 2008). The sequenced human genome was largely completed in 2001 (Venter *et al.*, 2001), cite Collins papers I & II), 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 (Levy *et al.*, 2007; Wheeler *et al.*, 2008), 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 individuals genome.

However, the cost of exhaustively collecting all of an individuals genetic information is currently too high to be done in the vast majority of studies. Instead, genetic variation 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 $\approx 0.1-1.0\%$ of the genome in most human populations. The vast majority of these have two allelic types.
2. Micro-satellites – regions of contiguous repeated nucleotide sequences where a sampled chromosome can have a different number of repeats. Each micro-satellite location thus often has more than two possible allelic types. Micro-satellites are currently thought to comprise $\approx 3\%$ of genetic variation in humans.
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 CNVs throughout the human genome span approximately 12% of the genome.

Most large-scale studies currently collect genetic information at $\approx 300-1000\text{K}$ 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 $\approx 0.1-0.2\%$ in the WTCCC study). Though markers consist only of 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. Furthermore, imputation algorithms (Marchini *et al.*, 2007) can be used to infer genetic variants that are not directly typed on an array or even to predict misclassified positions.

2.2 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 RNA transcripts (i.e. *pre-mRNA*) from DNA. The second stage, RNA processing, involves the splicing of *pre-mRNA* transcripts into mature RNA (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 joined back together 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

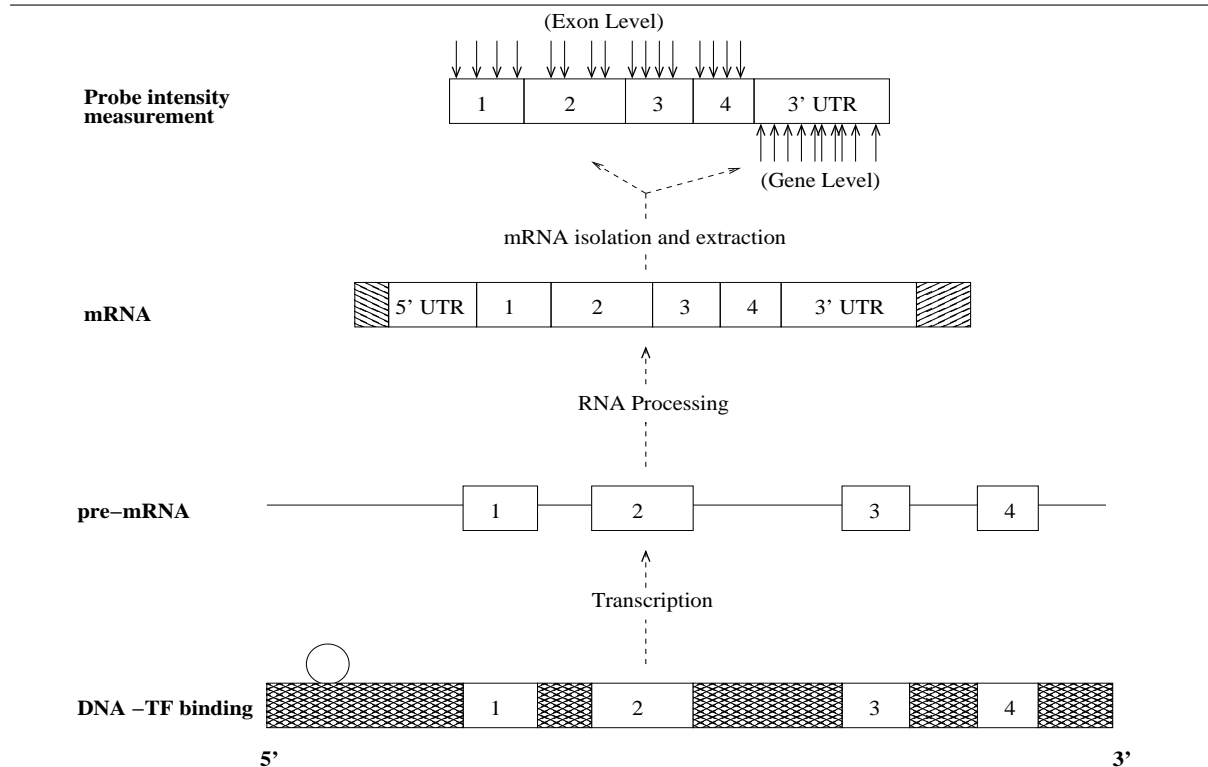
pre-mRNA, only the exons of genes may be transcribed into the mRNA that leaves the cell to synthesize proteins. It is therefore believed that the amount of protein synthesized by a gene can be interrogated indirectly by measuring mRNA levels in a cell (though mRNA levels do not necessarily correlate highly with protein levels – see section 4.3.1).

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 (Clamp *et al.*, 2007), whose mRNA transcript levels can now be routinely interrogated using a single “gene expression” microarray (table 1). 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” targeting a specific gene, though several probe sets may target the same gene. Microarrays designed to target multiple different exons throughout a gene can have upwards of 1 million probe sets each containing 4 probes. However, “standard” expression microarrays do not provide exon level data, with probe sets are located only towards the 3’ end of the gene. There are fewer probe sets in such “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 an 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. 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. They include micro-RNA (miRNA) and are different to regular protein coding RNAs in that they are shorter (\approx only 21-23 nucleotides) and complementary to mRNA. This enables them to bind to mRNA and 'interfer' with post-transcriptional processes and protein synthesis hence regulating protein gene expression. There are microarrays designed to target miRNAs specifically. There are approximately 800 known or putative human miRNAs which can be screened for simultaneously using the similar microarray technology.

2.3 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.

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, structural detection techniques are reserved for more focused studies and are not suitable for use on a global scale.

There are several types of technology which can target different proteins and/ or their structural properties (for example, phosphorylation states) and we describe the main resulting data types:

- 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-
- 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. The number of proteins probed for on a single array varies from x to y . (Hall *et al.*, 2007)

The authors of Chaerkady and Pandey 2008 provide an up to date review of these methods. Mention TF binding data (chip-ChIP/seq)?

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.

Something about QC? Are protein abundances also only relative or are the levels able to be interpreted biologically? What are the sources of noise? How much and what kind of samples can be profiled? What about variability between cells? Variability over time? Modelling dynamic behaviour? Spatial information?

2.4 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 the authors of Dunn and Ellis 2005. 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. cite:dunn,daouk for further details and discussion as to 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, in particular, they have shorter half lives (reference?) 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 couple 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.

2.5 Epigenomic Data

Epigenomics is the study of epigenetics at a global scale. There are a variety of definitions of epigenetics (Bird, 2007), 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 DNA packaging, in particular DNA methylation and histone modification.

Methylated DNA (mainly methylated cytosines) and histone modifications can be detected using chromatin immunoprecipitation techniques. There are specific antibodies against 5-methyl cytidine which can be used to immunoprecipitate methylated DNA fragments and similarly there are specific antibodies against various modified of 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. Bisulphate DNA can be hybridised to an array with probe pairs (one methylated (CG) and one unmethylated (CA)) surrounding CpG sites. 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 promotor 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.

2.6 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 (cite freimer and sabatti). Furthermore, there should be standardised procedures in place such that measurements are comparable across countries. In practise global large scaling phenotyping

in human individuals remains an ideal (with the human genome project proposed in 2003 (Freimer and Sabatti, 2003)) but there have been considerable advances towards this achieving this goal for mice (Bogue *et al.*, 2007). In particular, there are now centres throughout different countries where mice can be sent to be ‘phenotyped’ according to standardised tests.

In most studies human studies there are relatively few phenotypes catalogued which target morphological, behavioural and psychological traits. The majority are ‘clinical’ 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 of clinical phenotype characterisation has not improved in line with genetic, transcriptomic, proteomic and metabolic profiling and this mainly due to the subjective nature in which they are ascertained, for example different clinicians might ask different questions regarding symptoms of the disease. Consequently measurements/observations are often 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 cite Potsch08 *et al* call for supplementary phenotypes to distinguish between congenital heart disease conditions of different etiologies and cite bilder 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 phenotypes are expected as discussed previously, however, it is rare that multiple clinical 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 are assumed to 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 they 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’ (Kulzer and Orrit, 2004; Kou, Xie and Liu, 2005). 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 with 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. (What about limitations - restricted to model organisms cost? Could be seen as finer characterisation of phenotype but often impractical for the screening large numbers of individuals) Conclude with remark about external influences?

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
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. Range??
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 (what order of magnitude?)
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.	dimension??
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 three main concepts which we discuss in this section: a mapping from genotype to phenotype, networks and evolutionary relationships. 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. Finally, 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.

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). At the global level, entire genome to phenome mapping (equation 1) is rarely attempted. In this equation (and subsequent equations) $h(\cdot)$ and $f(\cdot)$ denote mapping functions and ε is noise is representative of all other factors not accounted for by the functions $h(\cdot)$ and $f(\cdot)$.

$$h(F) = f(G) + \varepsilon \quad (1)$$

G and F are high dimensional and establishing such a mapping would require the use of complex functions and highly computationally intensive modelling strategies. Furthermore as noted in section 2.6 high dimensional observations of the phenomes (excluding T, P and M) are not widely available for humans. Consequently, many studies simplify 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 2).

$$h(y) = f(G) + \varepsilon \quad (2)$$

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 3).

$$h(y) = f(g) + \varepsilon \quad (3)$$

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 4 where there are m loci affecting y).

$$h(y) = f(g_1, \dots, g_m) + \varepsilon \quad (4)$$

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 phenotype with probability 1 (equation 5). Note that equation 5 is an example of equation ?? where $h(y) = \mathbb{P}(y = 1)$ and $f(g_1)$ is the right hand term. Genetic effects are highly or completely penetrant for mendelian phenotypes such as sickle cell disease.

$$\mathbb{P}(y = 1) = \begin{cases} 1 & g_1 = g \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

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 6). In this equation e denotes epigenetic factors and x denotes external environmental exposures. These may also be multi-dimensional.

$$h(y) = f(g_1, g_2, \dots, g_m, e, x) + \varepsilon \quad (6)$$

The class of mapping functions in equation 6 is vast. The simplest approach is to consider the class of functions which describe main additive effects (equation 7).

$$f(g_1, g_2, \dots, g_m, e, x) = \sum_{i=1}^m f_i(g_i) + f_{m+1}(e) + f_{m+2}(x) \quad (7)$$

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 8 and 9 where there are separate terms for main effects and interactions.

$$f(g_1, g_2) = f_1(g_1) + f_2(g_2) + f_I(g_1, g_2) \quad (8)$$

$$f(g, e, x) = f_1(g) + f_2(e) + f_3(x) + f_4(g, e) + f_5(g, x) + f_6(x, e) \quad (9)$$

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. 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

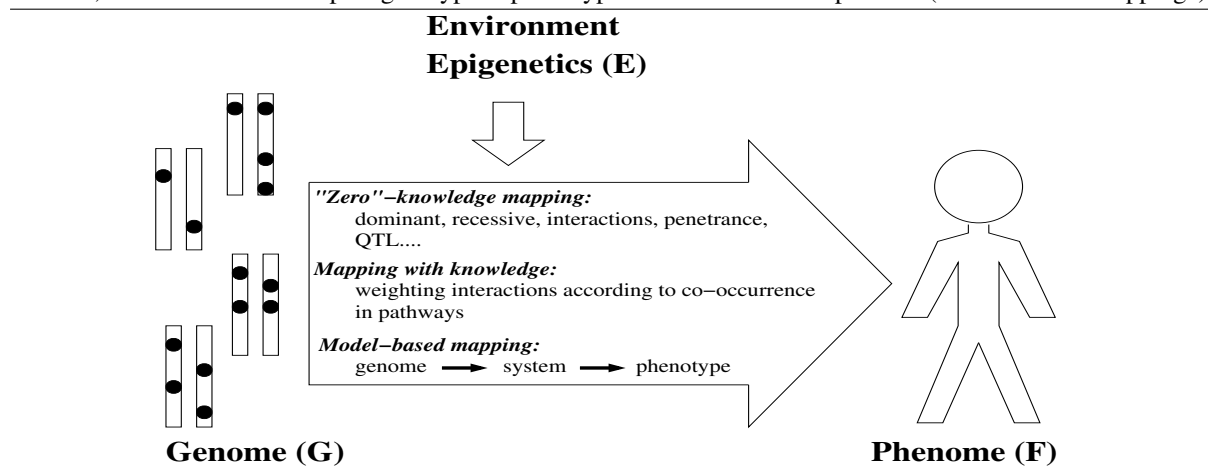
chi-squared contingency tables that test for independence between variants at a genetic locus g and case/control status (i.e. $y \in$ (“diseased”, “non-diseased”)) are often used to measure associations between genotype and phenotype (Balding, 2006). In a similar way genetic interactions influencing 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. (cite emily)

FUNCTION CLASS	FUNCTION	DESCRIPTION
Linear functions	$y = \alpha + \beta g + \varepsilon$	Single marker model
	$y = \alpha + \sum_{i=1}^m \beta_i g_i + \varepsilon$ $y = \alpha + \sum_{i=1}^m \beta_{1i} g_i + \sum_{i \neq j} \beta_{ij} g_i g_j + \varepsilon$	Multi-marker model 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
	$\log \left[\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} \right] = \alpha + \sum_{i=1}^m \beta_{1i} g_i + \sum_{i \neq j} \beta_{ij} g_i g_j$	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 notably disease case-control status.

Most functions including those described 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. It could be that increasingly ambitious modelling could reveal hidden components and lead to fundamental discoveries. Integrative approaches may well help in this regard, providing further knowledge to predict models that might be useful/appropriate for representing complex systems. If this will be the case remains to be seen.

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, then much more complex genotype to phenotype functions would be possible (“Model-based mapping”).



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?

3.2.1 Biological Networks

Biological networks are often used to describe physical systems, for example, a cell the size of E.coli with $\approx 10^9$ - 10^{10} molecules might be represented by a graph with say 10^3 - 10^4 nodes and edges according to the concentration of a subset of the molecules. Modelling human physical 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 $\approx 10^{41}$ 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 approximations (A1-5). It is not surprising therefore that resulting networks can rarely be representative of a complete physical system. How valid they are will only be revealed as increasingly ambitious attempts are made to simulate simple systems.

- A1 Biomolecules are approximated and represented by their observed concentration or number.
- A2 A large number of molecules do not feature in biological networks at all; either they are considered inessential or go unobserved.
- A3 Spatial effects are neglected.
- A4 Steady state assumptions apply to static observations of data such that biological statements can be made from observations of molecules at a single time point.
- A5 Dynamic behaviour of molecules can be approximated by a few characteristics such as rate parameters in a system of ordinary or stochastic differential equations.

Ideally biological network models should be dynamic to describe the temporal component inherent in all biological mechanisms. Furthermore, with a temporal component, the response to perturbations and/or intervention can be studied. In such instances A4 does not apply, however there are assumptions made about the time scale of the process, and establishing a sufficiently informative set of sampling times is difficult and in some cases harvesting data at this resolution is not possible.

There are four well established types of biological network (N1-4) 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.
- 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 hypergraphs rather than networks.

Notably, in each of these types, nodes generally represent the same type of biomolecule or at most feature two types of molecule (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 techniques used to reconstruct biological networks vary according to the types and sources of experimental data. Static and discrete temporal observations of a system are often analysed using statistical techniques (see the following subsection) and these can be effective for both small and large data sets. Dynamic modelling on a smaller scale (e.g. several interacting proteins, mRNAs or metabolites) can be done via the use of ordinary differential equations and continuous time stochastic processes. Dynamic modelling is particularly important for the reconstruction of signalling pathways, transcriptional regulatory networks and metabolic pathways.

3.2.2 Physical Networks

JOTUN TO WRITE A FEW LINES ABOUT THIS.

3.2.3 Statistical Networks

Statistical modeling is often employed to infer edges in networks. The resulting statistical networks differ from biological networks; firstly, 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 secondly, edges reflect dependence structure between nodes rather than functional relationships. Nodes of inferred statistical networks are dictated by the data types observed and the relationships between nodes are inferred via probabilistic modelling and/or the use of empirical correlation.

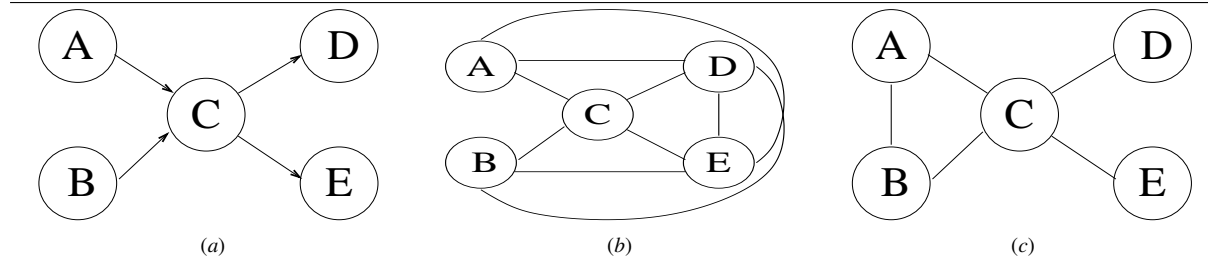
There are several important statistical principles and frameworks which are important and commonly used to infer statistical networks from biological data. We describe them below:

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. In the first instance nodes are connected if their empirical correlation exceeds a given threshold, where as the latter approach edges are weighted according to the strength of 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.
2. *Conditional Independence and 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. Complex dependence structures between variables can be represented by dependence graphs as we go on to describe or more generally with graphical models.
3. *Dependence Graphs*: Dependence graphs are undirected networks 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 expression data are usually undirected. Without additional information such as known transcription factors, genetic data or additional time series data direction of edges it is rare that

statements are made about the direction of the 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).

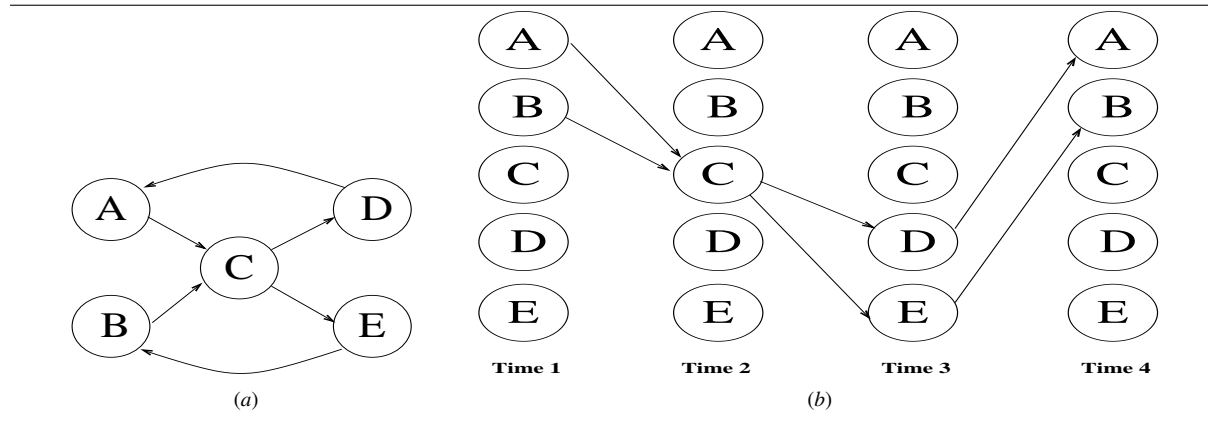
4. *Directed Graphs and Causality*: Adding directions to network edges provides a way of representing causality. A directed edge from node A to node B represents the fact that A influences B. For example SNP A might affect the expression of gene B. Directed edges can sometimes be inferred between nodes from different sources of data or by exploiting the flow of information which starts at DNA (according to the central dogma) (Zhu *et al.*, 2004), 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 and hence difficult to interpret. Directed Acyclic Graphs (DAGS) are a subset of directed graphical models in which there are no loops in the graph. They provide a convenient decomposition of the likelihood of the data and hence model fitting and parameter estimation is efficient consequently they are often used to reconstruct networks. It could be argued that DAGS are not suitable to represent biological networks since biological systems often contain feedback loops and complete cycles which are forbidden in DAGs. However, they can be modelled using Dynamic Bayesian Networks as described below.
5. *Dynamic (Bayesian) Networks*: Dynamic Bayesian Networks 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 (albeit via another node) then these variables will appear correlated. Correlation based networks might be useful to identifying this as a module of co-regulated genes amongst a larger set but it does not reflect the nature of the dependence structure. The dependence graph constructed on the basis of partial correlation 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 ?. 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 feed back loops can be modelled without creating directed cycles in the graph. This modelling strategy is however dependent upon the availability of temporal data.

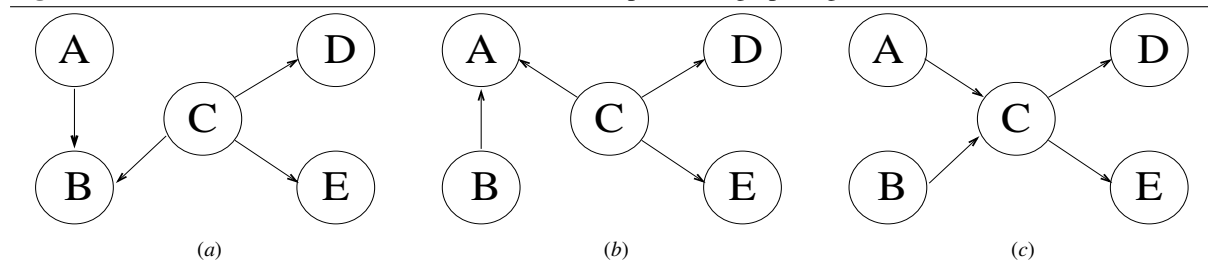
Figure 5 (a) A gene regulatory mechanism with a feedback loop. The system is as in figure 4 (a) with two additional feed back loops; Gene D regulates gene A and gene E regulates gene B **(b)** An example of how (a) can be represented with a directed acyclic graph and hence modelled using dynamic Bayesian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



3.2.4 Biological Interpretation of Statistical Networks

Biological interpretation of a statistical 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 networks consistent with statistical networks and further ambiguities arise when several statistical networks can 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, validate or generate new hypotheses. 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 highly

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

Alternative ways to interpret statistical networks biologically or to reduce the initial space of possible networks is to use existing biological knowledge (section 3.4) or observations from similar systems in related species. Exploiting data from other organisms relies on modelling of their evolutionary relationships as we describe in the following subsection.

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 follows and is usually observed using a subset of the genome. 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 cells reflect differences in state rather than differences in genetic material. At other levels such as regulatory or protein networks, shape, behaviour and more, 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 re-arrangements, CG-methylation, microsatellite repeat numbers. (It has for instance been calculated that each cell division is accompanied by about 50 errors in micro-satellites) and Copy Number Variation (CNV) and x sites of methylation. Thus knowing the genomic sequence for all cells would allow the recovery of the cell phylogeny

at high degree of certainty (Frumkin et al., 2005; Wasserstrom et al. 2008). 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

Individuals are related by a pedigree define by assigning parents to individuals. The pedigree for an individual are 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 (Helgasson) 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 were 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. Consequently the number of segments of ancestral material grows (and at most linearly) and furthermore, most ancestors further back in time, do not contribute genetic material to a specific individual such that the pedigree 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 have 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. In most data analysis the individuals are ignored. Treating individuals in a study as independent ignores the genetic dependence between individuals. Failing to account for this dependence can reduce both sensitivity and specificity of detecting genetic factors influencing phenotype. Individuals in a population are clearly related by a pedigree, but even if this information is partially known (e.g. as with family studies), as we stated previously, 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 (cite kingman, Hudson, HSW) and defining a probability measure on genealogical structures.

3.3.3 Genealogies relating species

There are in excess of 10^7 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 an other 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.

At present the main contributions of genome evolution fall in three categories; Firstly, observing the basic rates of events are of intrinsic value in understanding the mechanism of organismal change. Secondly, the strength and direction of selection. Identifying regions not under selection allows one to focus on regions of functional importance which is less than 10% in humans. There are more regions under purifying selection the positive selection (accelerating evolution relative to neutrality). Mapping negative (purifying) selection at the nucleotide or amino acid level is of great use in functional interpretation of the content of the genome. Position subject to accelerated change is of even great interest as they must underlie characters that must have changed for functional reasons in the specie history. Thirdly, the actual content of a genome can be interpreted functionally in terms of molecular mechanism.

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 ??), 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 which 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, is important to effectively utilise existing knowledge to minimise the number of false positive findings which are further investigated.

K1 Databases/Ontologies: These are structured and standardised vocabularies of concepts 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. ref.(Bodenreicher and Stevens, 2006; Bard and Rhee, 2004; Gkoutos et al., 2004) Mention Gene Cards?

- 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, 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 and Cardelli 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. (Calder et al., 2006; Baeten, 2005; Phillips et al., 2006; Aceto, 2004, Kwiatkowska et al., 2006)
- 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. (Raychaudhuri, 2006)

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 models itself - there is 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. The most famous example is hidden Markov models that now are ubiquitous and have very widespread applications in the biosciences. But it is a general principle and inherent in the way inference is done in biology.

The major examples of classes of models are:

Hidden Strings behind strings: Chomsky linguistic hierarchy from 1957 is famous for a variety of reasons. Chomsky grammars are originally deterministic and will be used to define by a finite set of rules, which string are belong to a language and which doesn't. If the use of different rules is assigned probabilities, then strings belonging the language will be assigned probabilities. If there is only one series of rule applications that can generate, then it is simple to assign a probability to this series. Two layers in this hierarchy - regular and context free grammars - have become widely used in biology. If their stochastic versions are used as hidden models, then we have hidden Markov models (HMM) and stochastic context free grammars (SCFG) respectively. Hidden Markov models have an enormous versatility. Major modelling areas in sequence analysis are: gene structures, protein secondary structures, signals and alignment. Stochastic context free grammars have especially been used to model RNA structures.

Hidden Processes 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 other cellular processes.

DOMAIN	OBSERVED	HIDDEN	REFERENCE
Haplotypes on a pedigree	Haplotypes	Inheritance	Lander and Green
Isochores in Genomes	Sequences	Isochores	Churchill
Sequence Alignment	Sequences	Alignment	Krogh
RNA Structure	Sequences	Structure	Durbin, Eddy, Hausler
Protein Secondary Structure (SS)	Sequences	SS	Goldman, Thorne and Jones
SS relationships	Sequences and SS	SS interactions	Abe and Mamitska
Protein Genes	a genome	gene structure	Burge and Karlin
Rate Variation	homologous sequences	fast/slow regions	Churchill and Felsenstein
Hidden Dynamic Processes	RNA levels	transcription factor	Lawrence and Rattray

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. GTH96 and CF are already comparative, while BK, DEH94 was made comparative in Pedersen and Hein (2003) and Knudsen and Hein (1999) respectively. C89, AM94 and LR07 could be combined with models of evolution to make comparative extensions with great benefit.

Modeling 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}. We focus on the analysis these data types collected from samples of human individuals for the study of phenotype. 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.

4.1 Analysis of single sources of data

4.1.1 Genomic Variation Data

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 it 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 1996 (Burge and Karlin) 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 (Ponting, 2008). The observation of conserved regions and other common features in the genomes 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. Firstly, a model of the distribution of the structure in question (e.g. gene structure) without knowledge of the nucleotide sequences. Secondly, 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 (e.g. HMMs reference) and RNA secondary structure (e.g SCFGs, reference). The two tier system can also be applied to annotate regulatory motifs although due to their short length, faster rate of insertion/deletion and the lack of simple characterising features it remains a challenge (Sandelin and Wasserman, 2004) 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 and 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. (SGJ?) 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 particularly 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 (examples?) although difficult to define are very useful in creating a common vocabulary for describing genes and biological processes. Furthermore, large international efforts HapMap (cite) and OMIM (cite) also catalogue variation at positions in the genome and variability/mutations associated with phenotypes.

I’m not sure how best to conclude this.

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. It appears that the vast majority of SNPs sampled from populations are rare, meaning that one of their alleles occurs at a small frequency in the population. For example, roughly half of SNPs have a *minor allele frequency*, i.e. frequency of the less common allele, less than 5% (is this true??). SNPs located physically near one another can be highly or perfectly correlated, though the level of linkage disequilibrium between them depends on the measure used and can also depend on their minor allele frequencies (Balding, 2006). Linkage disequilibrium between SNPs typically decreases exponentially as distance increases, and higher-than-average levels of linkage disequilibrium within human populations do not extend beyond 1-2Mb.

Model-based approaches applied to genetic variation data have also elucidated several important features of human biology. An example includes the discovery of recombination *hotspots*, or regions of 100 to 1000 times the recombination activity of regions outside of hotspots that act to break down linkage equilibrium between neighboring genetic locations (Myers *et al.*, 2005). Another example involves scanning the genome for regions 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 (Nielsen *et al* 2005). Comparisons between the three human populations of *HapMap* found evidence of selection in xx genes, including those involved in morphological characteristics such as skin pigmentation (Voight *et al* 2005).

A third example is the inference of the evolutionary history and levels of relatedness amongst humans. On average, humans are related by xx generations, and it is clear that genetic data varies detectably amongst human

populations. The genetic diversity of populations decreases as their geographic distance from Africa increases (Jakobsson *et al.*, 2008), 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. 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 (Novembre *et al.*, 2008). It is largely unknown whether these genetic differences have clear functional implications. However, they can be useful for mapping diseases as discussed below.

4.1.3 Molecular Phenotypes

As described in section 2, molecular profiling technologies can simultaneously quantify abundances of thousands of features. Structure and correlations within these quantities can be informative about the co-regulation and gene interactions. Clustering features according to similarities they exhibit across individuals is the most basic way. Clusters of features 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 5 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 (Weckwerth, 2003), it must be noted that these networks are distinct from protein interaction networks and metabolic reaction networks. These networks can only be inferred with 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> 1999
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 2005
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 2005, Friedman 2004
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> 2006, Segal <i>et al.</i> 2003, Ernst <i>et al.</i> 2007
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> 2007, Sanginetti <i>et al.</i> 2006

Table 5: 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. Examples include INSERT SOME EXAMPLES - PROCESSES IMPORTANT FOR DISEASE??

4.2 Analysis of phenotype with another source of data

Intro. Clarify definition of phenotype distinguishing between clinical and molecular.

4.2.1 Analysis with clinical phenotype with genetic data (G+F)

Detecting associations between genetic variants and disease susceptibility or variation in a 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. Studies of families: Genetic markers and clinical phenotypes are collected from families of closely related individuals (i.e. *pedigree data*) EXAMPLES Ceph families, deCode with SIZES
2. Studies of distantly related individuals: Genetic markers and clinical phenotypes are collected from 'independent' individuals sampled from a population (i.e. *population data*) EXAMPLES wtccc, SIZES.

Figure 7 illustrates the differences in the data resulting from these two sources and in this section we briefly outline principles underlying their analyses and limitations of these techniques.

Analysis of pedigree data

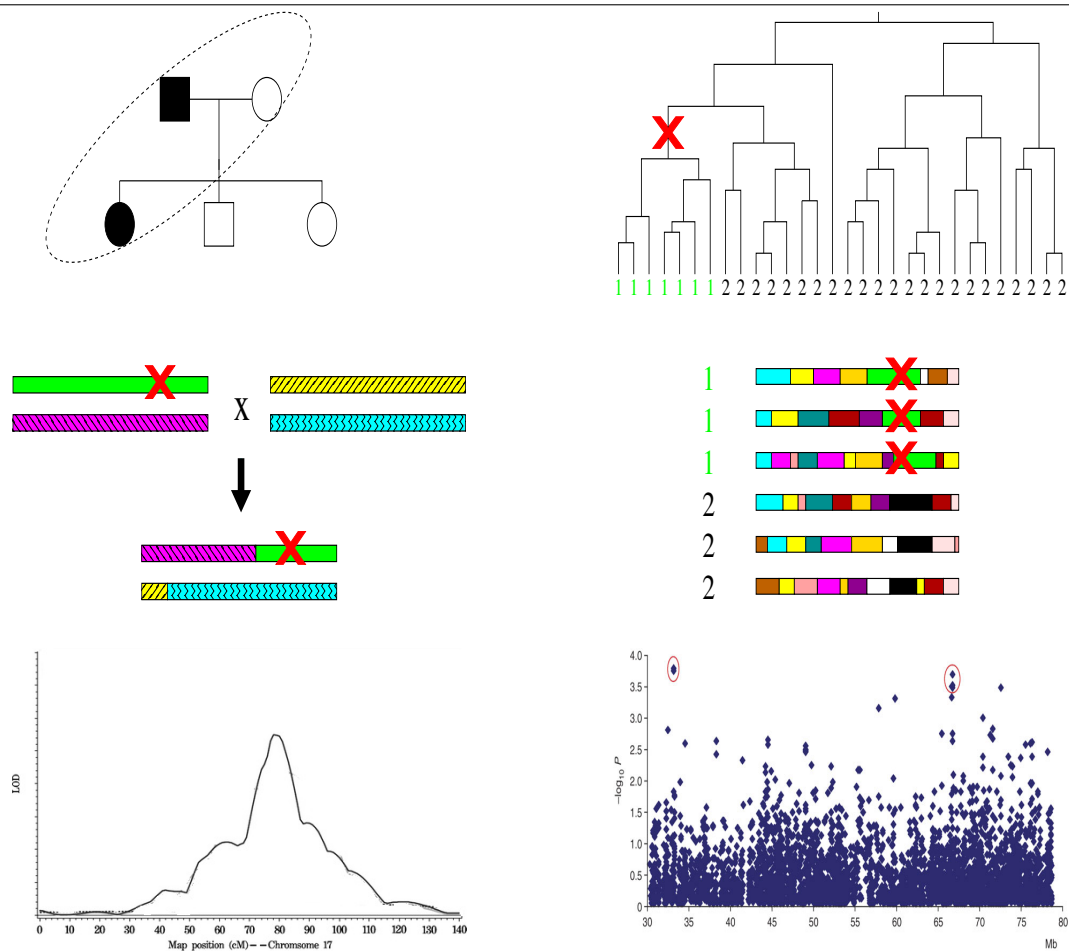
Genome-wide *linkage mapping* techniques elucidate genetic regions associated with a phenotype by studying the transmission of genetic material from generation to generation combined with the aggregation of phenotypes amongst family members in a pedigree. Key to this process is the DNA-shuffling biological phenomenon known as *meiotic recombination*. The net effect of a single recombination event on a chromosome in a single generation is that the DNA to the left of the event is inherited intact from one DNA source (or *haplotype*) of the previous generation, while the DNA to the right of the event is inherited intact from another DNA source of the previous generation. In regions of constant recombination rate (and ignoring selection and mutation), allelic types at basepairs of DNA located physically close together on a chromosome will more often be inherited together from a single DNA source than basepairs located physically far apart. Therefore a marker near a phenotype-influencing location will often be correlated with the alleles at that phenotype-influencing location, or *linked* to it. Recombination events are very rare (on average, only occurring about once per chromosome per generation), which enables the whole genome to be covered with relatively few markers. However, this also means that linkage mapping is typically only useful for identifying fairly broad regions harbouring phenotype-influencing location(s); such regions may be large (up to 10cM) and contain many genes. These concepts are illustrated in the left column of Fig.7.

Linkage techniques are most powered to detect highly penetrant single genes influencing a phenotype and have poor power to detect associations to complex multifactorial diseases since multiple genetic factors can each contribute small modifications to disease risk with potentially non-overlapping heterogeneous risk factors present across families in the same study. However, linkage mapping techniques have been able to identify genetic regions associated with Cystic Fibrosis (Kerem *et al.*, 1989), Huntington's disease, breast cancer, and muscular dystrophy. Some widely-used linkage mapping techniques include those described in Lander and Green 1987 and Olson *et al.* 1999. 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 2005 for a more detailed review).

Analysis of population data

Similar to linkage mapping, genome-wide association (GWA) studies attempt to detect associations between phenotypes and DNA variants. However, GWA studies do so using marker data collected in *unrelated* individuals, i.e. population data. Though the individuals in such datasets are unrelated at the family level, all individuals share evolutionary history. Therefore, as is the case with pedigree data, one might expect an individual to be more closely related in this history to an individual with a similar phenotype than to an individual with a wildly different phenotype, at least at the genetic location harbouring the phenotype-influencing variant (see Figure 7,

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 individuals father and the other inherited from the individuals 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 curvy 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 Lange et al 2007, Hum Genet 121:49-55. **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 Gudmundsson et al 2007, Nat Gen 39:977-983. Note how the score peaks are considerably more narrow compared to the linkage analysis results in the bottom left figure.



right). Furthermore, the expected length of shared contiguous DNA segments between individuals in such data is substantially lower than in familial data, because this history has been influenced by many recombination events that each act to break down associations amongst neighboring variants. As a consequence, GWA methods potentially allow for much greater resolution in the fine-mapping of phenotype-influencing loci. The most widely reported tests of association use *single marker association* 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 Balding (2006) for a detailed review).

SNPs are by far the most common employed markers in GWA studies, though micro-satellites and, more recently, CNVs have been used to detect associations using similar methods. Power to detect true associations via GWA methods depend primarily on three components:

1. Study sample size.
2. The effect size of the locus (often assessed by relative risk).
3. The minor allele frequency at associated genetic loci.

Associated with variants with small disease/phenotype allele frequencies and/or 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 (cite deCode, Broad, WTCCC, CGEMS, Cambridge, kingsmore08). 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.”

Some researchers have suggested that many diseases may be influenced by several such variants, perhaps each with individually small relative risks and/or minor allele frequencies. Still highly associated signals in narrow regions have been found and replicated in independent datasets for such diseases as prostate cancer, breast cancer, type II diabetes, ... etc...WHAT MORE DO WE WANT TO INSERT HERE?

4.2.2 Analysis with clinical phenotype and molecular phenotype (F + T),(F + P), (F + M)

These analyses ((F + T), (F + P), (F + M)) all aim to distinguish molecular phenotypes which correlate with the clinical 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 clinical phenotype.

Differential Gene Expression and Alternative Splicing 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, temporal changes in expression of homogeneous tissue. Determining which genes are significantly differentially expressed can be accomplished using a variety of scoring techniques. (WHICH ONES SHALL WE CITE?) Should we go into any more detail? Statistical issues - multiple testing, correlation of expression abundances?

Numerous disease studies over the past decade have exploited gene expression interrogation and analysis in applications for disease classification (Golub *et al.*, 1999; Sorlie *et al.*, 2003; Welsh *et al.*, 2001), prognosis (van 't Veer *et al.*, 2002) and drug response prediction (Chang *et al.*, 2003). A critical comparison and evaluation of microarray platforms (maq, 2006) 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 (Rakha *et al.*, 2008; Pusztai, 2008).

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 (Clark *et al.*, 2002) which involve computing a ratio of the abundances

of different 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 known to be associated and indeed causal to several diseases (Wang and Cooper, 2007) but with the recent availability of a global exon array, increasing numbers of studies seek to find differences in alternative splicing patterns which correlate with disease phenotypes (Thorsen *et al.*, 2008; ?; Soreq *et al.*, ???).

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.* (2008)) 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.

Limitations of gene expression profiling extend to biological interpretation of findings; causal and reactive gene expression perturbations 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 4.3.1.

Differential Protein Abundance/Structure

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 (Hanash *et al.*, 1986), heart disease (Mateos-Cáceres *et al.*, 2004), autoimmune diseases (Xiang and Kato, 2006) and Alzheimer's (Zhang *et al.*, 2005).

Some techniques used to quantify and characterise proteins (cite the review paper? and give 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. **example in here**. 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.

I feel like this needs a bit more maybe?

Differential Metabolite Abundance (Metabolic Signatures)

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". (cite examples - MNDs rozen05, Breast cancer Fan05, myocardial ischemia sabtine05) Furthermore, the composition of metabolic signatures (usually ≈ 10 distinct metabolites) can provide information about the processes perturbed by disease (cite another example?)

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 (Kaddurah-Daouk *et al.*, 2007).

4.2.3 Analysis with molecular phenotype with genetic data (G + T, P or M)

Discovery of QTLs for molecular phenotypes

One natural way of combining information from genetic and molecular phenotype data such as transcript or proteomic data 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. Currently considerable work has been done in this area for transcript abundance levels, i.e. gene expression. Cheung *et al.* (2003) provide evidence of familial aggregation of gene expression traits in human cell lines, and Schadt *et al.* (2003) suggest that 29% of the most variable gene expression traits expressed in Epstein-bar virus transformed lymphocytes are significantly heritable. Regions of the genome or individual loci found to be significantly associated with a gene expression trait are termed *expression quantitative trait loci* (eQTL). They are either cis-acting or trans-acting according to whether they reside on the same or different chromosomes to their expression target, respectively. This field of research has been termed “Genetical Genomics” following initial work by Jansen and Nap (2001).

Both genome-wide linkage (Morley *et al.*, 2004; Monks *et al.*, 2004) and genome-wide association studies (Cheung *et al.* (2005), Stranger *et al.* (2005)) have been used to search for eQTLs. The results of such studies are not entirely concordant, but as discussed by de Koning and Haley (2005) and Pastinen *et al.* (2006) 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 trait in some cases. Some caution must be taken, however, as large effects may be indicative of a cis-acting eQTL being in strong linkage disequilibrium with a SNP in a coding region of its target gene which affects the binding affinity of the mRNA rather than affecting the mRNA transcript abundances directly.

More recently Dixon *et al.* (2007); Goring *et al.* (2007) and Emilsson *et al.* (2008) each consider the distribution of the sizes of cis-acting effects. 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), they in general find that the most heritable traits have the strongest cis-acting effects. 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 $\approx 25\%$ whereas Dixon *et al.* (2007) 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.

Kwan *et al.* (2008) use exon tiling arrays to investigate the effect of SNPs on gene expression, thereby allowing the detection of genetic variants associated with alternatively spliced isoforms. They detect a total of 324 genes out of 17,897 which show expression or isoform association with one or more cis-acting SNPs, further characterising and quantifying the prevalence of the various different types of splicing events amongst the alternatively spliced isoforms. Their results suggest that genetic regulation of gene expression is more complex than previously shown with only 39% of detected genes exhibiting expression association with a nearby SNP across all exons, 29% showing association of transcription initialisation or termination and the remainder other alternative splicing events.

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. Notably, there are two recent studies cite holmes08 and melzer08 which report metabolite and protein quantitative trait loci respectively. CAN'T ACCESS HOLMES PAPER IN FULL. WOULD BE NICE TO COMMENT ON THEIR FINDINGS. The authors of Melzer *et al.* (2008) have shown that human protein abundance levels, measured using a variety of protein arrays, can similarly exhibit strong association with SNP data (Melzer *et al.*, 2008). Most of the reported SNPs with the strongest signals of association, termed *protein quantitative trait loci* (pQTL), were cis-acting. As the authors note, the understanding of the genetic influences underlying protein abundance likely have advantages over studying that underlying gene expression, as proteins are perhaps more directly involved in affecting the clinical phenotype. However, they only considered 42 proteins, perhaps a major reason why they failed to find any overlapping pQTLs among phenotypes as described in the next section for eQTLs.

In a similar way the effects of copy number variation on gene expression can be studied. 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.* (2007) suggest that the relative contribution to gene expression of CNVs to SNPs is small (17.7% : 83.6%), but with little overlap. 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. However as technology and ability to 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.

Directed networks regulatory networks

The central dogma dictating the flow of information from DNA to RNA can be exploited in the simultaneous analysis of genetic and transcriptomic data to construct directed gene regulatory networks. Zhu *et al.* (2004) are the first to do this and infer directed networks using a graphical models approach. They restrict model space by identifying potential parent nodes on the basis of a measure of genetic overlap which is calculated according to the strength of evidence for overlapping eQTLs. Although this approach integrates genetic data to make causal and regulatory statements, the resulting consensus network does not feature the eQTLs in the network directly, instead DAGs relating expression abundances are constructed.

Has anyone used genetic data with expression data and included SNPs in the graph?

4.3 Analysis with two molecular phenotypes (T+P, T+M, M+P)

The transcriptome, the proteome and metabolome are interactive, thereby motivating the simultaneous analysis of these data types. Protein coding RNA transcripts are translated to proteins and conversely, proteins regulate transcription of RNA. Furthermore, proteins interact with each other and with metabolites in functional cellular biochemical processes. final sentence??

4.3.1 Correlation between the Transcriptome and Proteome (T+P)

The highest correlation might be expected between the transcriptome and the proteome and indeed, transcript abundance is routinely quantified as an indirect measure of gene activity, however there is no consensus among scientists. Several studies investigate the correlation using a range of different samples ranging from yeast (Gygi *et al.*, 1999; Futcher *et al.*, 1999) through to human saliva (Hu *et al.*, 2006). Studies provide conflicting evidence; at a mere presence/absence level greenbaum suggest that upto 93% of 309 proteins found were also present as mRNAs where as at an abundance level, Gygi *et al.* (1999) suggest protein abundance vary widely from mRNA levels by upto 20 fold. Greenbaum *et al.* (2003) discuss reasons for possible lack of correlation and notably this includes the possibility that different profiling techniques target different expression regions of the gene and therefore are subject to confounding arising due to alternative splice variation (other possibilities?). A recent study (Bitton *et al.*, 2008) investigates this hypothesis by examining the correlation of the transcriptome and the proteome at an exon-peptide level. Reassuringly, they report improved exon-peptide correlations ($r^2 = 0.8$) relative to gene-protein correlations.

4.3.2 Correlation between the Transcriptome or Proteome and Metabolome (T+M, P+M)

There is no direct relationship between metabolites and gene transcripts or proteins so direct correlations are not expected between either the transcriptome or the proteome with the metabolome. Integrated analysis of metabolites with transcript and/or proteins can still provide insight to the biochemical processes and reactions which produce metabolites as their end products. Many 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. See Cuperlovic-Culf *et al.* (2008) for a review of integrated M+T and examples of applications of integrated M+T. P+T is less well characterised - cite lafaye et al 2005? should we go into more detail here?

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

Many multifactorial complex diseases 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 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 seems likely that it will further assist the identification of common sets of molecular pathways whose variation is associated with disease and aid the generation of causal phenotype/disease hypotheses.

There are two 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. (G + F with G + T, G + P, G + M)

One natural application of eQTL mapping is to search for overlap between eQTLs and (a) eQTLs identified from the expression patterns of other genes and (b) QTLs identified from other phenotypes. We use the term “overlapping” to mean that a region within a small window surrounding an eQTL also contains a locus or region associated with either another expression trait or phenotype. Again by exploiting the directional flow of information which starts at DNA, in some circumstances the discovery of overlapping eQTLs can suggest putative ways in which disease can manifest which would not be possible without drawing on both gene expression and genetic data together with a phenotype.

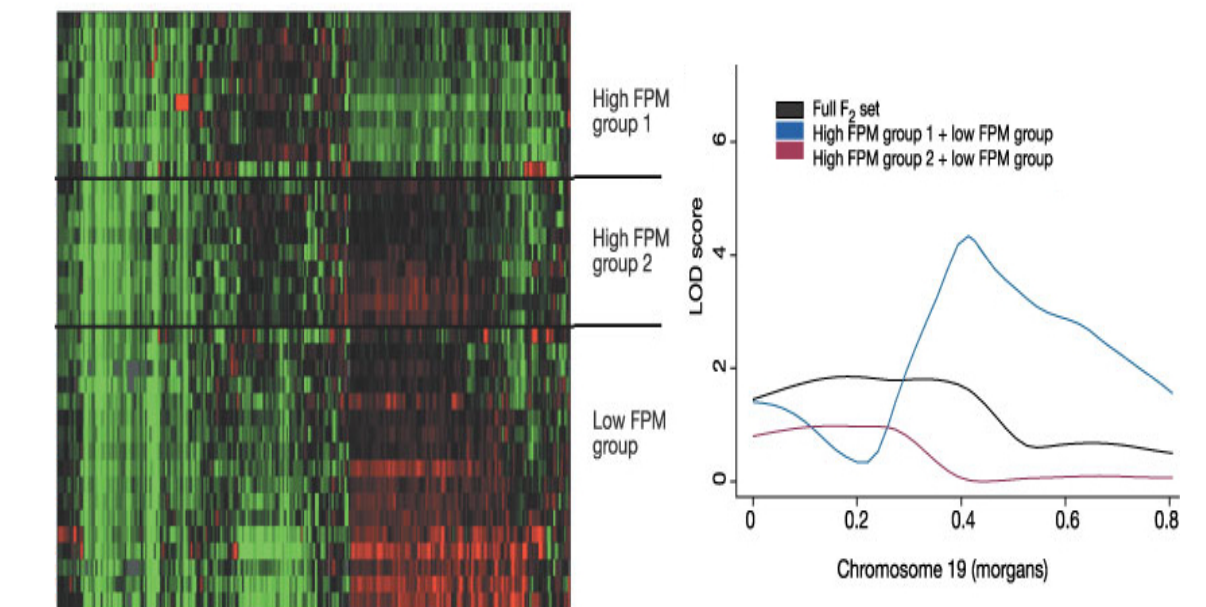
An example is the study of Moffatt *et al.* (2007). They report a highly significant genetic association to asthma on chromosome 17q23 spanning nineteen genes including gene ORMDL3. The same loci are also reported as eQTLs significantly associated with all transcripts of the ORMDL3 gene. Furthermore, 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.* (2007) and Sladek *et al.* (2007) combine their genetic association results with the global gene expression study of Dixon *et al.* (2007) and discover/prioritise candidate genetic regions for Crohn’s disease and Type-2 diabetes, respectively. In another example, Goring *et al.* (2007) find an eQTL for a gene which overlaps with a locus associated with a quantitative clinical phenotype which itself highly correlates with Cardiovascular disease thereby supporting the proposition of this gene as a candidate for modifying disease risk via the quantitative clinical phenotype.

There are few (or no?) 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 important 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 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 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.

Figure 8 Results of a F_2 mouse cross from Schadt *et al.* (2003), depicting a QTL analysis based on sub-phenotypes (right) that were selected based on the clustering of expression values (left). For the expression plot, genes go left to right and the different mice samples go top to bottom. The different expression patterns across genes were used to cluster the mice with high fat pad mass (FPM) into two separate groups, one of which showed a strong signal of association on chromosome 19 (blue line in the right plot).



4.4.2 Integrated Networks

As previously discussed in section 3.2 network models provide a flexible framework from which multiple data sources can be analysed and findings represented. Although they are founded on huge assumptions, integrated networks in which edges are inferred between and within different data types can be used 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 $3^{\binom{n}{2}}$ for directed networks and $2^{\binom{n}{2}}$ for undirected networks (where n is the number of node in the graph) so fitting each model 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.

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 is more extensive and validated in model organisms relative to human and that the underlying systems are less complex. Model organism examples (Tu *et al.*, 2006; Sun *et al.*, 2007) are founded on extensive biological knowledge, where as Schadt *et al.* (2005) restrict inference to testing selected pathways on the basis of overlapping eQTLs. tu use both approaches and sample pathways from a network constructed using existing biological knowledge.

This needs a bit more I think.

4.5 Analysis of all types of biological data 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 of humans.

The principle uses of model organisms are summarised in table 6.

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 and more 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 (ref) 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.

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 6: 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 Focused Studies

6 Conclusion

In this review, integrative genomics has major components (data, concepts, knowledge, models and analyses), data had a few major groups. These are clearly coarse categories that in any study can be complemented by a variety of data types and information that enters in an intuitive informal way. However, it is our hope that the components and data types still does describe major features of present research and as such is useful in extracting some unity in what can seem bewilderingly complex.

Major goals of the biosciences are clearly full understanding and finally predictive modelling of biological systems. This will have a major bottom up component. The genome wide studies are describing systems of a size that cannot be modelled in the foreseeable future. Genome wide studies have been tremendously useful in coarse functional assignments of genes, but cannot lead to detailed understanding of biological systems. This will have to come from more detailed understanding of smaller systems or fewer genes. The goals of integrative genomics are to give a statistical model of the observed quantities that allows causal inference. Doing this is hard as correlations can be weak and the data very noisy. And if this wasn't bad enough, the goals of the integrative biology are much weaker than the goals of systems biology that attempts a physical model of biological systems that can be used for prediction and control. Systems biology is so ambitious that this could well be many years into the future before it is achieved. However, it is the natural next step as each time a causal connection is established attempting for a physical explanation will be undertaken. How hard this is and for how large systems it can be done, is impossible to predict.

Directly or indirectly, most of the funding for the biosciences are motivated by the hope of alleviating human disease. The overall trajectory of research will be Mapping → Region → functional interpretation → control/drug development. The first two steps are well described and at “functional interpretation”, integrative genomics or systems biology offers to be the tools of choice. However, these techniques are still too high level, with concepts such as “protein”. Functional interpretation aims at a detailed molecular story that potentially could go to an atomic description. The techniques applied here is mainly laboratory experiments will only small role of the statistics used in integrative genomics. The final step “control/drug development” is a classic field needing huge resources. It seems that functional interpretation is the really hard step that will only do progress by large of amounts of high quality data focused on few relevant components complemented by traditional biochemical analysis. Modeling will be as relevant as ever. There will still be a need to transfer knowledge from model systems to typically humans and evaluate the reliability. But the models will have to describe how small systems are interpreted in detail. A long series of large scale genome wide studies have come to a grinding halt at proving a functional interpretation and is forced to have that as the next top priority.

Integrative Genomics and most of Systems Biology use observations related to the cell and information and data involving higher levels play a small if any role, with the exception of phenotype. Given the fundamental impossibility of calculating the genotype → phenotype function presently and in the foreseeable future, all observations that help characterising this function are of value. Many new data types, such as medical imaging, relating to tissues and organs will be of increasing relevance.

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. Slightly more than 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. Another present trend that will continue is the relevance of external knowledge to any experiment. This is already occurring in the form of public databases, but is being complemented by increasing advanced and flexible methods to represent and manipulate knowledge. Examples of this are ontologies (ref), the semantic web (ref) and text mining (ref).

References

- (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotech*, **24**(9), 1151–1161.
- Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**(10), 781–791.
- Bird, A. (2007). Perceptions of epigenetics. *Nature*, **447**(7143), 396–398.
- 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+.
- Bogue, M. A., Grubb, S. C., Maddatu, T. P., and Bult, C. J. (2007). Mouse Phenome Database (MPD). *Nucleic acids research*, **35**(Database issue).
- 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).
- Chaerkady, R. and Pandey, A. (2008). Applications of proteomics to lab diagnosis. *Annual review of pathology*, **3**, 485–498.
- 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.
- 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.
- 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.
- 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+.
- 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.
- Cuperlovic-Culf, M., Belacel, N., and Culf, A. (2008). Integrated analysis of transcriptomics and metabolomics profiles. *Expert Opinion on Medical Diagnostics*, pages 497–509.
- Dawn and Barrett, J. H. (2005). Genetic linkage studies. *The Lancet*, **366**(9490), 1036–1044.
- de Koning, D. J. and Haley, C. S. (2005). Genetical genomics in humans and model organisms. *Trends Genet*, **21**(7), 377–381.
- 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.
- Dunn, W. B. and Ellis, D. (2005). Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry*, **24**(4), 285–294.
- 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., Helgadóttir, A., Jonasdóttir, A., Jonasdóttir, A., Styrkarsdóttir, U., Gretarsdóttir, S., Magnusson, K. P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H. G., Stefansson, T., Leifsson, B. G., Thorsteinsdóttir, 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*.
- Ernst, J., Vainas, O., Harbison, C. T., Simon, I., and Bar-Joseph, Z. (2007). Reconstructing dynamic regulatory maps. *Mol Syst Biol*, **3**.

- Freimer, N. and Sabatti, C. (2003). The Human Phenome Project. *Nat Genet*, **34**(1), 15–21.
- Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, **303**(5659), 799–805.
- 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.
- 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.
- 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.
- 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).
- 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.
- Hall, D. A., Ptacek, J., and Snyder, M. (2007). Protein microarray technology. *Mechanisms of ageing and development*, **128**(1), 161–167.
- 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.
- 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.
- 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.
- Jansen, R. C. and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics*, **17**(7), 388–391.
- 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).
- 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.
- 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*.
- 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.
- 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+.
- 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).
- 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*.
- 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.
- 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., Ruokonen, 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).
- 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*.
- 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.
- 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.
- 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.
- 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*.
- Olson, J. M., Witte, J. S., and Elston, R. C. (1999). Genetic mapping of complex traits. *Statistics in medicine*, **18**(21), 2961–2981.
- Pastinen, T., Ge, B., and Hudson, T. J. (2006). Influence of human genome polymorphism on gene expression. *Hum Mol Genet*, **15 Spec No 1**.
- Ponting, C. P. P. (2008). The functional repertoires of metazoan genomes. *Nature reviews. Genetics*.
- Pusztai, L. (2008). Current Status of Prognostic Profiling in Breast Cancer. *Oncologist*, **13**(4), 350–360.
- 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.
- 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.
- 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.
- 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.

- Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754–764.
- 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**.
- 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*.
- Soreq, L., Gilboa-Geffen, A., Berrih-Aknin, S., Lacoste, P., Darvasi, A., Soreq, E., Bergman, H., and Soreq, H. (2007). [identifying alternative hyper-splicing signatures in mg-thymoma by exon arrays. *PLoS ONE*, (6).
- 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., Lippman, 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.
- 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).
- 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.
- Sun, W., Yu, T., and Li, K.-C. (2007). Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics*, **23**(17), 2290–2297.
- 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.
- 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+.
- 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).
- 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., Kerkhoven, 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.
- 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., Yooseph, 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.
- 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.
- Weckwerth, W. (2003). Metabolomics in systems biology. *Annu Rev Plant Biol*, **54**, 669–689.
- 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.
- Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., Mcguire, A., He, W., Chen, Y.-J., Makhijani, 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.
- Xiang, Y. and Kato, T. (2006). Use of proteomics in analysis of autoimmune diseases. *Lupus*, **15**(7), 431–435.
- Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, **4**(1).
- 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).
- 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.