1 An introduction to systems genetics

Florian Markowetz and Michael Boutros

Systems genetics is an emerging field based on old approaches going back to the genetic studies performed by Gregor Mendel (Mendel 1866). Mendel's experiments primarily focused on explaining inheritance of single traits and their phenotypes – for example how specific genetic alleles influence colour or size of peas – but recently developed technologies can comprehensively dissect the genetic architecture of complex traits and quantify how genes interact to shape phenotypes by using natural variation or experimental perturbations as a basis to understand links from genotypes to phenotypes. This exciting new area has recently been termed 'systems genetics' (Civelek & Lusis 2014).

While the basic, underlying questions are not new, systems genetics builds upon major methodological advances that facilitate the measurement of genotypes and phenotypes in a previously unforeseen and comprehensive manner. With this arsenal at hand, one of the major aims of systems genetics is to understand "how genetic information is integrated, coordinated and ultimately transmitted through molecular, cellular and physiological networks to enable the higher-order functions and emergent properties of biological systems" (Nadeau & Dudley 2011).

1.1 Definition of systems genetics

Systems genetics is born out of a synthesis of multiple fields: it integrates approaches of genetics, genomics, systems biology and 'phenomics', that is, our increased ability to obtain quantitative and detailed measurements on a broad spectrum of phenotypes. One of the first papers using the term 'systems genetics' defines it as "the integration and anchoring of multi-dimensional data-types to underlying genetic variation" (Threadgill 2006). Since then, many studies have aimed at integrating genome-wide data across many different levels, and possibly different environments, in approaches that are closely related to quantitative genetics.

In our view, a systems genetic approach should bring together three dimensions: it should combine (i) a genome-wide analysis with (ii) many quantitative phenotypes, both at the molecular and organismal level, (iii) in many different conditions or environments (Fig. 1.1). As such, the integration that needs to be achieved by systems genetics goes

Systems Genetics: Linking Genotypes and Phenotypes, ed. F. Markowetz and M. Boutros. Published by Cambridge University Press.

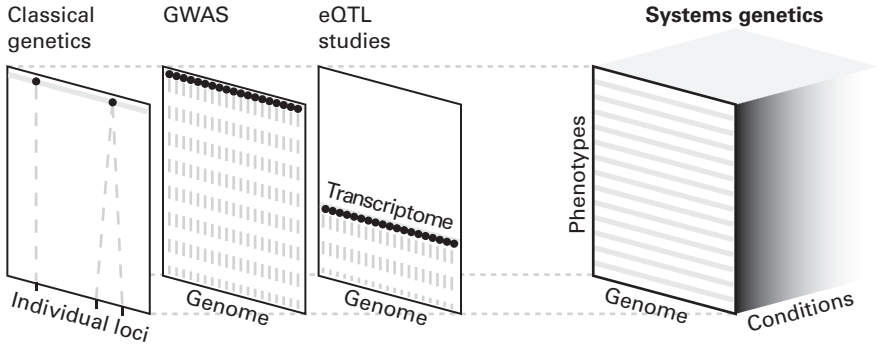

Figure 1.1 Systems genetics ideally combines genome-wide analysis with many quantitative phenotypes, both at the molecular and organismal level, in many different conditions or environments. This subsumes previous approaches that were focused on linking individual loci to a single phenotype (in classical genetical and epistatic analysis) or comprehensively linking many genomic loci to a single phenotype in a single condition (in GWAS or eQTL studies; see text).

well beyond a 'single' dimension of genotypes or phenotypes. For example, genome-wide association studies (GWAS) link natural variation to a single phenotype in a single condition. While systems biology often focuses on the dynamic behaviour of systems, systems genetics uses genomic techniques to link genotypes with large-scale, quantitative measurements of phenotypes.

While in general this might sound like a lofty goal which is hard to achieve, studies that come close already exist in model systems like plants (Sozzani & Benfey 2011, Topp et al. 2013), *Drosophila* (Ayroles et al. 2009) or human cell lines (ENCODE Consortium et al. 2012). For example, single-nucleotide polymorphisms (eQTL) analysis uses expression quantitative trait loci (SNPs) and transcripts to establish links between genotypes and phenotypes. The strength of such approaches is most easily demonstrated in model organisms where large amounts of phenotype and genotype data at different levels already exist. For example Skelly et al. (2013) measured transcript, protein, metabolite and morphological traits in 22 genetically diverse strains of *Saccharomyces cerevisiae* to gain insights into the spectrum and structure of phenotypic diversity and the characteristics influencing the ability to accurately predict phenotypes. Similarly in *Drosophila*, Ayroles et al. (2009) integrate gene expression and phenotypes across multiple conditions, leading to the identification of overlapping groups of transcripts that influence different organismal phenotypes (see Fig. 1.2). These studies address many fundamental questions, such as how many different SNPs influence the same phenotype (complex phenotypes) or how many different phenotypes are influenced by the same SNP (pleiotropy). Because such studies look globally across many phenotypes, they can shed light on interactions between genes. Genetic interactions, also called epistatic interactions, can then be used to dissect the genetic architecture of particular processes (Avery & Wasserman 1992, Phillips 1998, Phillips 2008).

Why is this an important area of research? Firstly, many traits are complex and not determined by a single gene. Instead, combinations of alleles with low abundance or

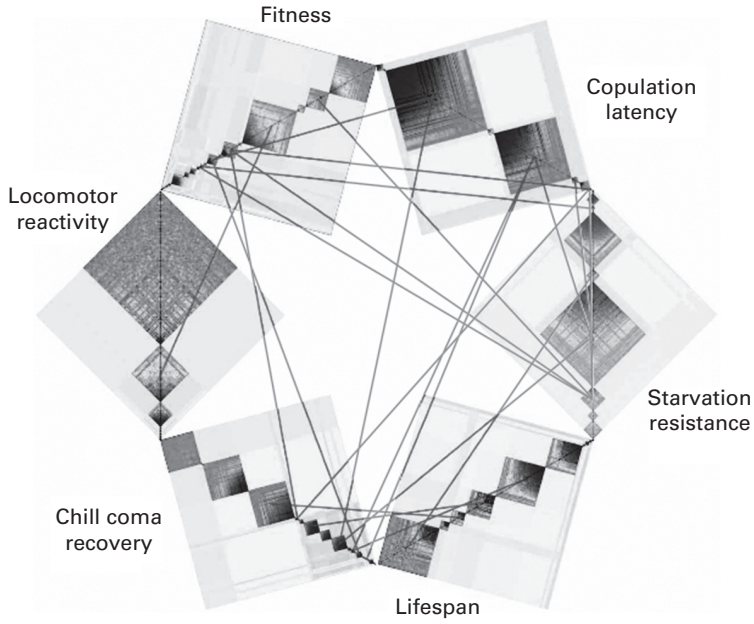

Figure 1.2 Pleiotropy between phenotypic modules in *Drosophila*. Grey lines connect modules with a significant overlap of greater than four genes. Adapted by permission from Macmillan Publishers Ltd (Ayroles et al. 2009). A black and white version of this figure will appear in some formats. For the colour version, please refer to the plate section.

context-dependent effects (environmental or genetic background) are responsible for the bulk of phenotypic variation. Secondly, genetic interactions have also been proposed to account for missing heritability (Zuk et al. 2012) but this issue is still contentious and widely discussed (Eichler et al. 2010, Hemani et al. 2013). Thirdly, comprehensive network models built on principles of correlation and causation might be able to predict phenotypic outcome in response to different genetic backgrounds and environmental or therapeutic perturbations. These models could also identify targets for modulating phenotypic outcome which would be an important contribution to treatment and prevention of disease. These technological advances have been argued to entirely transform healthcare in the future (Friend & Ideker 2011).

1.2 History of systems genetics

The simplest example of a systems genetic question is how two genes influence a single phenotype. This question and the concept of epistasis have a long tradition in genetics, with theoretical underpinnings that go back more than a hundred years.

Epistasis: from one gene to two genes

Different definitions of epistasis emerged that vary between sub-disciplines (Phillips 1998, Cordell 2002, Phillips 2008). Bateson's original definition is based on the idea of

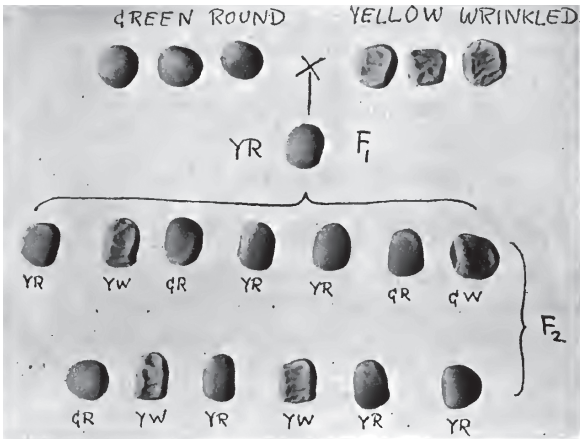

Figure 1.3 Seed-character inheritance after Bateson (1909).

one gene 'standing above' another (Greek: epistasis) (Bateson 1909) and is sometimes called 'compositional epistasis' or 'masking' (Phillips 2008).

Bateson wrote on epistatic relationships: "Pending a more precise knowledge of the nature of this relationship it will be enough to regard those factors which prevent others from manifesting their effects as higher, and the concealed factors as lower. In accordance with this suggestion the terms epistatic and hypostatic may conveniently be introduced" (Bateson 1909, p79) (see Fig. 1.3).

Fisher's statistical definition of epistasis (originally called 'epistacy') compares the phenotype of a two-loci genotype with the phenotype predicted by a model simply adding up the effects of the two loci (Fisher 1918). If the prediction of the additive model differs from the observed phenotype, the two loci are called epistatic (Phillips 2008). These concepts were established a hundred years ago, decades before genetic information was understood to be encoded by DNA.

Genomics: from one gene to genome-wide

In the last few decades the amount of genomic information has dramatically increased. The first eukaryotic genome, *S. cerevisiae*, was sequenced in 1995 and since the early 2000s the human genome sequence has been determined and the collection of human genes is known. Microarray and later sequencing technologies have led to the identification of genetic variants and the ability to measure gene activity genome-wide. New experimental tools like gene deletion libraries, RNA interference (RNAi) and transposon mutagenesis allow manipulation of all the genes in a genome (Boutros & Ahringer 2008). In observational studies, epistasis can be measured in genome-wide association studies (GWAS) (Cordell 2009), which allows modelling of genetic traits on a genome-wide level. Perturbation studies, combining loss-of-function alleles to assess their combined phenotypes, have been spurred by genome-wide deletion libraries. For example in yeast, a large part of the 18 million possible genetic interactions has been discovered by systematic double-knockout experiments (Costanzo et al. 2010)

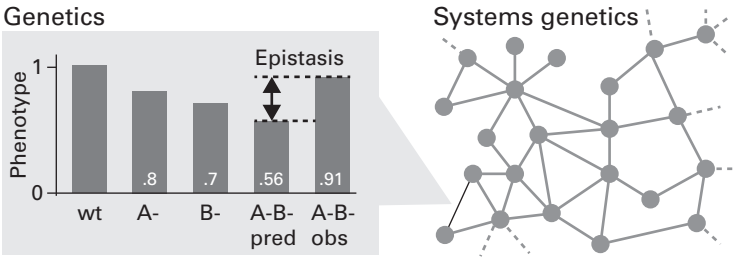

Figure 1.4 Systems genetics comprehensively uses concepts from classical genetics. One example is epistasis: where the phenotype of a double perturbation deviates from the prediction based on the phenotypes of the single perturbations. Automated, large-scale approaches extend this analysis to a genome-wide level (Costanzo et al. 2010). Each edge in a large genetic interaction network corresponds to one classical genetics experiment.

(see Fig. 1.4). Similarly, genetic epistasis has been discussed as a major driver underlying genotype-to-phenotype relationships of complex traits and of genome evolution (Mackay 2014).

Phenomics: from one phenotype to many phenotypes

In the past years, there has been much progress in technologies to quantitatively measure phenotypic data on a genome-wide scale. Molecular phenotypes, such as expression levels or transcription factor binding sites have been mapped in many organisms and across many conditions. For example, to better understand how genetic variation influences gene expression levels, genetic linkage and association mapping have identified cis- and trans-acting DNA variants in so-called expression quantitative trait loci (eQTL) studies (Cheung & Spielman 2009). The ENCODE (and modENCODE) projects generated many of such phenotypes that can be correlated to underlying genetic or environmental changes, including regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome sequence (Celniker et al. 2009, ENCODE Consortium et al. 2012). At one level above, phenotypes can be measured at the level of the cell. For example, large-scale studies have interrogated the whole genome for genes that are required for cell proliferation (Berns et al. 2004, Boutros et al. 2004, Cheung et al. 2011, Marcotte et al. 2012, Vizeacoumar et al. 2013). Similarly, cells can be interrogated for more complex phenotypes, such as changes in cell shape, cell division, cell migration or to the response of specific signalling pathways (Kiger et al. 2003, Snijder et al. 2009, Fuchs et al. 2010, Snijder et al. 2012, Yin et al. 2013). Whole-organism phenotyping, e.g. in plants, has become feasible through novel technologies that can comprehensively characterize phenotypes in an automated manner (Sozzani & Benfey 2011). Phenome-wide association studies can help to understand the genetic basis of diseases: "There is great interest in making use of the large amounts of phenotypic data that are stored in electronic medical records (EMRs) in the quest to understand the genetic basis of disease. One way to do this is to carry out phenome-wide association studies (PheWASs), in which genetic variants are tested for association with a wide range of phenotypes" (Flintoft 2014).

Systems genetics: a synthesis of genetics, genomics and phenomics

Modern systems genetics differs in many important aspects from historic genetic approaches. Modern approaches can rely on quantitive and often automated measurements of phenotypes, which can be now be achieved on individuals (cells, organisms) rather than on populations. Genotyping has become feasible for hundreds or even thousands of individuals, and within an individual, deep sequencing allows to identify subsets of genetically different cells, e.g. by assessing genetic heterogeneity in tumours.

Advances in biological and medical imaging by high-resolution microscopy or functional MRI allow to measure phenotypes in a much more quantitative manner. Also, technological advances allow measurements of a variety of intermediate phenotypes in a high-throughput manner, including chromatin state, histone modifications and gene and soon protein expression. The ENCODE is a good example of a large-scale phenotyping project that characterised large parts of the genome (ENCODE Consortium et al. 2012). The analysis of these comprehensive data sets is enabled by new computational approaches to represent and interpret biological information. Examples are the network and graph theory approaches that have permeated wide parts of biological research in the last few years (Barabási & Oltvai 2004, Ideker & Krogan 2012). Network models can help to understand and interpret the roles of genetics and epigenetics in disease predisposition and etiology. By providing the backbone of molecular interactions through which signals are transduced and gene expression is regulated, networks have been proposed to limit the search space of allele variants and alterations that can be causally linked to the presentation of a phenotype (Califano et al. 2012).

The combination of new high-throughput experimental technologies with new computational analysis methods allows a comprehensive view on complex interactions between genes, beyond measuring epistasis on individual gene pairs. The focus of research has shifted from analysing individual genetic interactions to comparing detailed profiles of genetic interactions, which are much more informative of cellular genetic architecture (Costanzo et al. 2010, Baryshnikova et al. 2013). These technologies are far beyond what we were able to envision in the year 2000 and promise to revolutionise our understanding of the genetics of complex traits.

In cancer, genomic efforts to characterise tumour genomes on many molecular levels (The Cancer Genome Atlas Network 2012) have started to be complemented with quantitative assessment of cellular and tissue morphology (Beck et al. 2011, Yuan et al. 2012). Cell line studies have led to a better mechanistic understanding of how cancer genes influence complex cellular phenotypes, such as cell morphology and invasion (Yin et al. 2013). In pathology, computational methods reach back half a century (Smith & Melton 1964) and modern methods are mainly being used for quantifying immunohistochemical stainings in tissue microarrays (Schüffler et al. 2010, Fuchs & Buhmann 2011, Schüffler et al. 2013). Computational methods have been very successful in making standard pathological analyses more objective and reproducible, but they have generally not quantified the global spatial organisation of the tumour tissue and were generally not linked closely to genomics. With large multi-level data collections, e.g in breast cancer (Curtis et al. 2012, The Cancer Genome Atlas Network 2012), this situation is beginning to change (Ali et al. 2013) and we expect it to lead to a systems genetics understanding of

cancer that links tumour genotypes to intermediate molecular and tissue phenotypes as well as organismal phenotypes like progression, survival and treatment response (Yuan et al. 2012).

1.3 Future challenges

We see a number of areas where we believe systems genetics will have an immediate impact. One such area is systems genetics of disease, where some work has already been done in metabolic syndromes (Schadt et al. 2005), but less so in cancer research. Looking at cancer from a systems genetics perspective offers new opportunities for understanding basic concepts, finding ways to apply combination treatments and for the development of new therapeutic strategies.

Tumours are ideal objects for systems genetic analyses, because they contain many different genotypes, many different cellular phenotypes, ongoing clonal evolution, interactions with the organismal environment and between the cells within the tumour microenvironment. These interactions influence disease progression and outcome as well as the development of metastases and are of key importance for treatment decisions. Often treatments fail because of outgrowth of a resistant cancer subpopulation. Systems genetics approaches will have great impact on understanding how heterogeneity in cancer genotypes influences different outcomes. Systems genetics approaches might also lead to rational design on combinatorial drug treatment, identifying key intervention points that might be masked by buffering through other pathways.

Cancer is also an ideal application for systems genetics because many genomes and intermediate phenotypes have already been collected, on a population scale in projects like The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) and also by multiple sampling within individual patients. However, the question of how to connect genotype to phenotype on a single-cell level in a complex tissue is still an experimental and theoretical open problem. An experiment one could envision to tackle this question is to take one metastasis, determine the genome of every single cell and as phenotypes the spatial organisation of the tumour tissue and molecular phenotypes like gene expression. This would allow to associate individual genotypes with the environment the cell lives in and link them to intermediate molecular phenotypes as well as morphology. This will depend on further advances in single-cell genomics techniques, which is currently a rapidly emerging research area (Shapiro et al. 2013).

Another area where we predict a large impact of sytems genetics is in complementing observational studies with large-scale perturbation studies that build on foundations in yeast and extend them to more complex organisms. First examples exist in flies, worms and human cells (Ayroles et al. 2009, Muellner et al. 2011). Understanding the genetic architecture of organisms more complex than yeast is an important ongoing area of research. Additionally, how drug treatments interact with the genotypes of individual cells to influence subpopulations and lead to resistance is an important field.

The success of these research programmes will depend on theoretical and computational advances. Standards for the basic data analysis of single-cell phenotyping have

yet to be established. In addition, how to integrate data from different molecular and phenotypic levels in populations of cells is currently poorly understood.

1.4 What is covered in the book

Systems genetics is a broad field with many aspects ranging from population genetics to molecular networks, which cannot all be covered in a single volume. In this book we focus on how experimental perturbations can help us to understand the link between genotype and phenotype. We provide a snapshot of current research activity and state-of-the-art approaches to systems genetics.

The book chapters cover both experimental and theoretical approaches: on the experimental side the topics range from large-scale RNA interference studies and mutant analysis to combinatorial perturbations and epistatic interactions; on the theoretical side from network reconstruction and reliability analysis to data integration and conceptual discussion of the nature of phenotypes. To show the wide applicability of systems genetics methods, we chose work from a variety of model organisms, including *S. cerevisiae*, *Drosophila melanogaster* and human.

Myers and colleagues (Chapter 2) introduce the definition of genetic interactions and describe how comprehensive genetic interaction analysis has been performed in yeast.

Boutros and colleagues (Chapter 3) introduce genetic interaction analysis and how it has been applied to map genetic networks in different organisms. They introduce how genetic interactions can be experimentally measured and how quantitative genetic interaction profiles can be deduced in order to make conclusions about the genetic architecture of processes.

Beerenwinkel and Delgado-Eckert (Chapter 4) concentrate on the mathematical analysis of genetic interaction networks. They introduce a mathematical framework to link epistasis to the redundancy and reliability of biological networks. They present statistical methodology for analysing epistatic relationships.

Whitehurst and Maxfield (Chapter 5) extend the discussion of genetic interactions and epistasis by describing how to use functional screening approaches to identify genetic vulnerabilities in cancer cells. Overcoming chemoresistancy and exploiting chemosensitivity is an important area of cancer research and this chapter presents how synthetic lethal analysis can make its way from the bench to the bedside.

Markowetz and colleagues (Chapter 6) give an overview of statistical analysis strategies for genetic screens ranging from genome-wide screens with single reporters to targeted screens with rich molecular phenotypes. They describe statistical methods for functional annotation and network reconstruction.

Meyer and colleagues (Chapter 7) describe the application of genetic perturbation studies in infectious diseases, and how RNAi screens identify novel host cell targets with previously unknown roles in infection and its pathology.

Perrimon and colleagues (Chapter 8) expand on the description of network inference methods by focusing on the integration of gene perturbation studies with mass

spectrometry data, which can play an important role in generating hypotheses, driving further experimentation and providing novel insights.

Girolami and colleagues (Chapter 9) describe an application of Bayesian model selection to signalling pathways in chronic myeloid leukaemia. They describe how gene perturbations can be integrated with a dynamic phenotype to infer pathway structure.

Bader and Park (Chapter 10) describe statistical modelling of dynamic protein complexes and apply them to networks in *S. cerevisiae* and *Arabidopsis thaliana*.

Linding and Hadjiprocopis (Chapter 11) take a broader view on phenotypes and different phenotype states and describe theoretical concepts to formalise cellular decision-making.

In the last chapter **Schafer and Brown** (Chapter 12) go beyond cells to organisms and behavioural phenotypes, how they can be quantitatively measured in *Caenorhabditis elegans* and linked to phenotypes.

References

Ali, H. R., Irwin, M., Morris, L., Dawson, S.-J., Blows, F. M. et al. (2013), 'Astronomical algorithms for automated analysis of tissue protein expression in breast cancer.' *Br J Cancer* **108**(3), 602–612.

Avery, L. & Wasserman, S. (1992), 'Ordering gene function: the interpretation of epistasis in regulatory hierarchies.' *Trends Genet* **8**(9), 312–316.

Ayroles, J. F., Carbone, M. A., Stone, E. A., Jordan, K. W., Lyman, R. F. et al. (2009), 'Systems genetics of complex traits in *Drosophila melanogaster*.' *Nat Genet* **41**(3), 299–307.

Barabási, A.-L. & Oltvai, Z. N. (2004), 'Network biology: understanding the cell's functional organization.' *Nat Rev Genet* **5**(2), 101–113.

Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B. & Boone, C. (2013), 'Genetic interaction networks: toward an understanding of heritability.' *Annu Rev Genomics Hum Genet* **14**, 111–133.

Bateson, W. (1909), *Mendel's principles of heredity*, Cambridge University Press.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O. et al. (2011), 'Systematic analysis of breast cancer morphology uncovers stromal features associated with survival.' *Sci Transl Med* **3**(108), 108–113.

Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A. et al. (2004), 'A large-scale RNAi screen in human cells identifies new components of the p53 pathway.' *Nature* **428**(6981), 431–437.

Boutros, M. & Ahringer, J. (2008), 'The art and design of genetic screens: RNA interference.' *Nat Rev Genet* **9**(7), 554–566.

Boutros, M., Kiger, A. A., Armknecht, S., Kerr, K., Hild, M. et al. (2004), 'Genome-wide RNAi analysis of growth and viability in *Drosophila* cells.' *Science* **303**(5659), 832–835.

Califano, A., Butte, A. J., Friend, S., Ideker, T. & Schadt, E. (2012), 'Leveraging models of cell regulation and GWAS data in integrative network-based association studies.' *Nat Genet* **44**(8), 841–847.

Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S. et al. (2009), 'Unlocking the secrets of the genome.' *Nature* **459**(7249), 927–930.

Cheung, H. W., Cowley, G. S., Weir, B. A., Boehm, J. S., Rusin, S. et al. (2011), 'Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer.' *Proc Natl Acad Sci USA* **108**(30), 12 372–12 377.

Cheung, V. G. & Spielman, R. S. (2009), 'Genetics of human gene expression: mapping DNA variants that influence gene expression.' *Nat Rev Genet* **10**(9), 595–604.

Civelek, M. & Lusis, A. J. (2014), 'Systems genetics approaches to understand complex traits.' *Nat Rev Genet* **15**(1), 34–48.

Cordell, H. J. (2002), 'Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.' *Hum Mol Genet* **11**(20), 2463–2468.

Cordell, H. J. (2009), 'Detecting gene–gene interactions that underlie human diseases.' *Nat Rev Genet* **10**(6), 392–404.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. et al. (2010), 'The genetic landscape of a cell.' *Science* **327**(5964), 425.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M. et al. (2012), 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.' *Nature* **486**(7403), 346–352.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M. et al. (2010), 'Missing heritability and strategies for finding the underlying causes of complex disease.' *Nat Rev Genet* **11**(6), 446–450.

ENCODE Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D. et al. (2012), 'An integrated encyclopedia of DNA elements in the human genome.' *Nature* **489**(7414), 57–74.

Fisher, R. A. (1918), 'The correlations between relatives on the supposition of Mendelian inheritance.' *Trans R Soc Edinburgh* **52**, 399–433.

Flintoft, L. (2014), 'Disease genetics: phenome-wide association studies go large.' *Nat Rev Genet* **15**(1), 2.

Friend, S. H. & Ideker, T. (2011), 'Point: are we prepared for the future doctor visit?' *Nat Biotechnol* **29**(3), 215–218.

Fuchs, F., Pau, G., Kranz, D., Sklyar, O., Budjan, C. et al. (2010), 'Clustering phenotype populations by genome-wide RNAi and multiparametric imaging.' *Mol Syst Biol* **6**, 370.

Fuchs, T. J. & Buhmann, J. M. (2011), 'Computational pathology: challenges and promises for tissue analysis.' *Comput Med Imaging Graph* **35**(7–8), 515–530.

Hemani, G., Knott, S. & Haley, C. (2013), 'An evolutionary perspective on epistasis and the missing heritability.' *PLoS Genet* **9**(2), e1003295.

Ideker, T. & Krogan, N. J. (2012), 'Differential network biology.' *Mol Syst Biol* **8**, 565.

Kiger, A. A., Baum, B., Jones, S., Jones, M. R., Coulson, A. et al. (2003), 'A functional genomic analysis of cell morphology using RNA interference.' *J Biol* **2**(4), 27.

Mackay, T. F. C. (2014), 'Epistasis and quantitative traits: using model organisms to study gene–gene interactions.' *Nat Rev Genet* **15**(1), 22–33.

Marcotte, R., Brown, K. R., Suarez, F., Sayad, A., Karamboulas, K. et al. (2012), 'Essential gene profiles in breast, pancreatic, and ovarian cancer cells.' *Cancer Discov* **2**(2), 172–189.

Mendel, G. (1866), 'Versuche über Pflanzen-Hybriden.' *Verhandl Naturforsch Vereines Brünn* **4**, 3–47.

Muellner, M. K., Uras, I. Z., Gapp, B. V., Kerzendorfer, C., Smida, M. et al. (2011), 'A chemical–genetic screen reveals a mechanism of resistance to PI3K inhibitors in cancer.' *Nat Chem Biol* **7**(11), 787–793.

Nadeau, J. H. & Dudley, A. M. (2011), 'Genetics: systems genetics.' *Science* **331**(6020), 1015–1016.

Phillips, P. C. (1998), 'The language of gene interaction.' *Genetics* **149**(3), 1167–1171.