
CHAPTER 17

Concluding Remarks: Final Thoughts and Future Trends

MICHAEL R. BARNES¹ and IAN C. GRAY²

¹*Genetic Bioinformatics and* ²*Discovery Genetics
Genetics Research Division
GlaxoSmithKline Pharmaceuticals, Harlow, Essex, UK*

- 17.1 How many genes?
 - 17.2 Mapping the genome and gaining a view of the full depth of human variation
 - 17.3 Holistic analysis of complex traits
 - 17.4 A final word on bioinformatics
 - Acknowledgements
 - References
-

The sequencing of the human genome is complete. This is an obvious milestone for all fields of biology, none more so than genetics. As we have seen throughout this book, the availability of a complete genome makes the study of the genetics of an organism much less haphazard, and bioinformatics is an essential enabling skill for the geneticist to make the most of the genome. In the pre-genome era, geneticists probed the genome like early explorers penetrating a dark continent ripe for exploration. Relying on only the most basic data they painstakingly reconstructed genes and methodically drafted maps to find disease alleles. Now, in the post-genome era, instead of stars and a compass the genetic explorers have the equivalent of a global positioning satellite system and a detailed A–Z directory of genes. With all this technology it might be hard to imagine how genetics can now fail to locate disease genes, but failure will still be a frequent outcome.

Why? Firstly, we may be looking for something that does not exist or is too small to detect using existing methodology. Complex human disease is a product of both environment and genes, but the environment is often overlooked as a source of disease, particularly in the current era of high-profile genetics. The contribution of a single gene to a multifactorial disease or trait may be vanishingly small and consequently even large studies may have insufficient power to detect it. Secondly our directory of genes may not be as comprehensive as we think, with significant weaknesses in certain areas, for example the assignment of function to poorly understood regulatory motifs and the degree and nature of inter-individual genome diversity. Thirdly our maps are not yet completely

error free. Bioinformatics cannot help with the first problem directly, although novel statistical methods may improve the chances of identifying small genetic effects and will form part of a continually evolving software suite for genetic analysis of complex traits. As more pieces of the genetic and environmental jigsaw puzzle are put into place for each complex trait, it should become progressively easier to position the remaining pieces to give a more complete picture. Although bioinformatics may be perceived as playing a secondary role in developing techniques for improved statistical analysis of complex trait data, it is the key to providing the equally important solutions required for a truly complete characterization of the genome coupled with unimpeachable data integrity.

17.1 HOW MANY GENES?

The biggest revelation of the human genome sequencing project was that humans appear to have fewer genes than we had expected. Estimates of the total number of human genes were widely anticipated to reach the 100,000 gene mark (Aparicio, 2000). As sequencing progressed these estimates were downgraded to 60–70,000 and finally as the first draft appeared estimates were consolidated to a mere 35,000 genes (Ewing and Green, 2000). If this figure is to be believed, then humans have only seven times as many genes as yeast, ~2.5 times as many as the fly *Drosophila melanogaster* and less than twice as many as the nematode worm *Caenorhabditis elegans*. This figure may increase as understanding of the genome and gene prediction increases, although it seems unlikely that the number will rise beyond 50,000.

This smaller than expected number of genes might be viewed as good news for geneticists—fewer genes to screen for disease association. But fewer genes does not necessarily equate to reduced complexity. Complexity can manifest at many levels, including splicing, gene regulation, post-transcriptional editing and post-translational modification. In Chapter 12, we described the *Drosophila* DSCAM gene which has 115 exons which are alternatively spliced to code for 38,016 related but distinct protein isoforms (Schmucker *et al.*, 2000). This remarkable gene gives us a hint that many of the gene models described so far in humans could under-represent the true diversity of the human gene repertoire. Instead it may be wise to view every gene transcript as a unit specific to a particular tissue, time or cellular condition. Alterations in any of these conditions could direct the expression of an alternative transcript.

It may also be pertinent to question the definition of a gene. Traditionally a gene is viewed as a protein-coding unit. Transcripts which do not obviously code for a protein are often dismissed as ‘regulatory RNA’—a virtual dumping ground for transcripts which we are just beginning to understand (see Szymanski and Barciszewski, 2002). This situation is exacerbated by the wealth of data generated by genomics; for example a very large number of ESTs and cDNAs show no *in silico* evidence of splicing (i.e. by each end aligning either side of an intron in a genomic sequence). There are a number of explanations for the existence of such transcripts. They could be derived from a real gene but simply do not span an intron and therefore show no evidence of splicing; alternatively they could be *in vitro* artefacts generated during the construction of cDNA libraries or *in vivo* artefacts generated from cryptic promoters or pseudogenes.

This highlights one of the biggest challenges for the bioinformatic interpretation of the human genome—data overload. Gene prediction and annotation tools generally disregard unspliced ESTs as supporting evidence for the existence of a gene. This is a necessary precaution to avoid over-prediction of genes across the genome; tools designed to analyse whole genomes have to sacrifice sensitivity to avoid extensive over-prediction of genes

and to maintain the performance of genome analysis pipelines, but where geneticists seek to identify all candidate genes in a defined locus, it may be prudent to evaluate equivocal information such as unspliced ESTs in a more thorough fashion. This can be achieved easily with genome browser tools such as Ensembl and the UCSC human genome browser which present all available data across a locus. However, it is wise to proceed with caution when planning experimental work based on ambiguous data derived from *in silico* sources in order to avoid frustration as well as wasted time and resources. Simple, rapidly executed experiments to provide supporting evidence for the *in silico* observation should be the first step.

17.2 MAPPING THE GENOME AND GAINING A VIEW OF THE FULL DEPTH OF HUMAN VARIATION

Our incomplete understanding of genes and genome organization may not necessarily be a big problem for genetics. Experimental frameworks can be primarily focused on the physical and genetic composition of a region, in terms of genetic markers, recombination frequency and other characteristics, rather than its perceived functional content. 'Phenotype-driven' family-based whole genome linkage scans to identify genes responsible for monogenic traits illustrate one such approach. Use of linkage disequilibrium (LD) to identify genomic regions of genetic association is a second example, and is more appropriate for complex traits. This approach assumes little about the function of a marker or gene, but can allow mapping of a genetic association to a very small region (typically 10–100 kb) following the construction of detailed population-based LD maps. Completion of an LD map of the entire human genome will in itself be a highly significant milestone for genetics. Already provisional LD maps of chromosomes 21, 22 and 19 have been published (Dawson *et al.*, 2002; Patil *et al.*, 2001; Michael Phillips personal communication). A whole genome LD map generated by many of the former members of TSC should be made publicly available in late 2003. This will finally make comprehensive SNP-based whole genome association scans a realistic possibility; selecting SNPs which tag all of the major haplotype blocks across the genome will shift the emphasis toward good experimental design and away from conjecture when initiating genetic association studies.

However, evolution toward a whole-genome haplotype-based approach to genetic studies will present considerable challenges. For example, although all of the available evidence suggests that the majority of haplotypes in any given genomic region are common to multiple ethnic groups (Gabriel *et al.*, 2002), haplotype frequencies may vary considerably between groups. Thus markers that tag common haplotypes in one ethnic group may not identify the most common haplotypes in other groups. Furthermore, approaches based on attempts to associate common haplotypes with a disease state are broadly reliant on the veracity of the 'common disease caused by common variants' hypothesis (see Pritchard, 2001). A low frequency haplotype which is associated with disease may evade detection, and a rare predisposing SNP occurring on a common haplotypic background may not be detected due to insufficient statistical power. Only empirical data gathered over the next few years will reveal the true scale of such issues. A further consideration is the increase in throughput and reduction in cost required to render the necessary scale of genotyping for population-based association studies, which are likely to require several million data points per genome-wide experiment, feasible. However significant investment in this area has led to promising improvements across a range of genotyping platforms over the last few years and we expect this trend to continue.

17.3 HOLISTIC ANALYSIS OF COMPLEX TRAITS

One of the weaknesses of genetic association studies is the difficulty in drawing a firm conclusion regarding the robustness of the finding from the statistical evidence for association between a given gene and trait, particularly if the level of significance is marginal. A key future application of bioinformatics is likely to be the drawing together of diverse threads of data from a number of sources in a more holistic approach toward the analysis of complex traits. The output from human linkage and population-based association studies can be combined with animal model quantitative trait loci, phenotypic data from systematic gene knock-out and transgenic mouse approaches, genome-wide expression data from microarrays, proteomic profiles and other sources, to provide a substantial body of evidence relating to the gene or locus in question. This will require the development of both new interfaces for the integration of disparate datasets and sophisticated global analysis software.

17.4 A FINAL WORD ON BIOINFORMATICS

It is always difficult to present a rapidly moving field such as bioinformatics in a book. Despite the best efforts of the authors, editors and publisher, by the time this book reaches the reader many of the tools described in the preceding chapters will have evolved to offer yet more functionality and utility. Keeping abreast of new developments in bioinformatics is as important an activity as using the data themselves. Current awareness of the field is essential to ensure that all of the relevant available data are captured, maximizing research efficiency. Finally, the best approach to becoming proficient in the use of software tools is often trial and error, and bioinformatics is no exception; trial and error *in silico* can obviate the far less desirable prospect of trial and error in the laboratory, so do not be afraid to experiment with bioinformatics applications — see what the human genome can yield in your hands. Good luck!

ACKNOWLEDGEMENTS

MRB and ICG would like to acknowledge the efforts of all of the authors who have contributed to this volume. This book has taken shape after many discussions with many of our colleagues at GSK and in the wider scientific community. The first drafts were moulded into final chapters with the assistance of several willing proof readers, particularly Christopher Southan, Aruna Bansal, Ralph McGinnis and Mary Plumpton. We would also like to express our gratitude to Joan Marsh, Layla Paggett, Amie Tibble and Monica Twine at John Wiley for able assistance in the preparation of the manuscript. Finally this volume would not have been possible without the support and encouragement of Robin Dement and Ian Purvis at GSK.

REFERENCES

- Aparicio SA. (2000). How to count human genes. *Nature Genet* **25**: 129–130.
- Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, *et al.* (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544–548.

- Ewing B, Green P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet* **25**: 232–234.
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, *et al.* (2002). The structure of haplotype blocks in the human genome. *Science* **296**: 2225–2229.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
- Pritchard JK. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**: 124–137.
- Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, *et al.* (2000). Drosophila DSCAM is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671–684.
- Szymanski M, Barciszewski J. (2002). Beyond the proteome: non-coding regulatory RNAs. *Genome Biol* **3** (reviews 5): 1–8.