ESSAY

The road to genome-wide association studies

Leonid Kruglyak

Abstract | The recent crop of results from genome-wide association studies might seem like a sudden development. However, this blooming follows a long germination period during which the necessary concepts, resources and techniques were developed and assembled. Here, I look back at how the necessary pieces fell into place, focusing on the less well-chronicled days before the launch of the HapMap project, and speculate about future developments.

Genome-wide association studies (GWAS) use dense maps of SNPs that cover the human genome to look for allele-frequency differences between cases (patients with a specific disease or individuals with a certain trait) and controls. A significant frequency difference is taken to indicate that the corresponding region of the genome contains functional DNA-sequence variants that influence the disease or trait in question. The recent crop of results from GWAS (reviewed in REFS 1–4) might seem like a sudden development. However, this blooming follows a long germination period during which the necessary concepts, resources and techniques were developed and assembled. Here, I look back at how the necessary pieces fell into place, starting with early ideas and continuing with concrete proposals and theoretical and empirical studies that laid the foundation for The International HapMap Project[5]. I close by contemplating the implications of the lessons that were learnt from the initial crop of GWAS for future studies of human genetic variation.

Early milestones

Genome-wide approaches to human genetics date back to the proposal in 1980 by Botstein and colleagues for the construction of a linkage map of the human genome, with restriction fragment length polymorphisms (RFLPs) as molecular markers[6,7]. The natural initial applications were to genetically simple Mendelian diseases; however, as early as 1986, even before the first linkage map was completed, Lander and Botstein recognized that most human traits and diseases follow complex modes of inheritance, and they discussed several approaches for studying such complex traits[8]. One of the approaches they proposed was linkage disequilibrium (LD) mapping, which recognizes that a mutation that is shared by affected individuals through common descent will be surrounded by shared alleles at nearby loci, representing the haplotype of the ancestral chromosome on which the mutation first occurred (FIG. 1). The first example of LD between a DNA polymorphism and a disease mutation was provided by an association between an allele of an RFLP in the β-globin gene and the sickle-cell form of haemoglobin[9].

> “[Genetic] complexity is present on multiple levels, and might be fruitfully thought of as ‘fractal’.”

Simple population-genetics arguments suggested that LD in the general human population would probably be limited to distances below 100 kb[10]. For this reason, Lander and Botstein deemed LD mapping to be impractical in the general population owing to the high marker density that would be required, but they proposed that a map of hundreds of RFLPs might suffice for LD mapping in recently founded isolated populations.

The first complete RFLP map of the human genome was reported in 1987 (REF. 11), but human mapping studies really flourished once microsatellites replaced RFLPs[12] (BOX 1). Genome-wide studies used family-linkage approaches almost exclusively, with LD being used to refine the locations of genes that were mapped by linkage, as pioneered by Kerem and colleagues for the cystic fibrosis gene[13]. In a groundbreaking and forward-looking study in 1994, Houwen and colleagues reported the first application of LD mapping in an initial whole-genome search for a disease locus[14]. Following the approach that was envisioned by Lander and Botstein almost a decade earlier, they used 256 markers to map the gene responsible for benign recurrent intrahepatic cholestasis (BRIC) in an isolated fishing community in the Netherlands. Their success relied on the rarity of the disease and on the availability of a population isolate in which the affected individuals were distant relatives. A similar study in a Mennonite kindred allowed Puffenberger and colleagues to identify a gene for Hirschsprung disease[15]. However, such studies, which straddled the border between family linkage and LD mapping, remained the exception as linkage approaches dominated. Studies of many pairs of relatives (most commonly, affected sib pairs) were especially prevalent owing to the ease of collecting such samples versus samples from extended families, to their theoretical appeal for mapping complex traits[16–18] and to the availability of powerful analysis tools[19]. These genome scans were carried out for many common diseases that show complex inheritance, but they failed to find many reproducible loci. With these findings, the initial belief that a few major genes would explain susceptibility to complex diseases gave way to the realization that the level of complexity was much

higher and that many loci of individually small effect were involved. Because such loci are difficult to identify by family linkage owing to limited power, the search was on for new approaches.

Modern proposals

In an influential perspective, Risch and Merikangas argued that association studies should be more powerful than family linkage studies for detecting high-frequency, small-effect polymorphisms[20]. Linkage studies rely on allele sharing by descent among affected relatives, and their low power to detect such polymorphisms is due to two factors. First, when the increased risk conferred by an allele is small, some relatives will be affected because of other causes and will not carry the risk allele. Second, when an allele is common, it can enter the family through multiple founders, erasing clear inheritance patterns. These effects combine to decrease sharing by descent to the point at which an impractically large number of families must be studied to detect it. Association studies still suffer from the first effect (which is inherent to searches for small effects) but not the second, and therefore they have a higher sensitivity in detecting common variants with small effects. The increase in power is such that even testing large numbers of polymorphisms, with the ensuing statistical costs of multiple testing, does not erase the advantage of association studies[20,21]. In addition to offering higher power with the same sample sizes, association studies also have the practical advantage that large samples of unrelated cases and controls can be collected much more easily than family-based samples.

Risch and Merikangas issued a call for a catalogue of all variants in human genes, and set out a challenge "to the molecular technologists to develop the tools" for their identification and genotyping[20]. This call was echoed by Eric Lander, who hypothesized that common variants of modest effect might hold the key to susceptibility to common diseases[21] (this was subsequently codified as the common disease–common variant hypothesis). Lander also noted that the role of noncoding variation might be studied by the use of LD mapping with a sufficiently dense polymorphism map.

These proposals were formalized the following year in a policy forum by Collins, Guyer and Chakravarti[22]. They made explicit the distinction between the direct approach of cataloguing all common functional variants and the indirect approach of relying on a dense map of SNPs for LD mapping (FIG. 2). A back-of-the-envelope calculation put the likely size of shared ancestral haplotypes in the range of 10–100 kb, leading to a proposal to identify at least 100,000 SNPs. To achieve this goal, The SNP Consortium, a public–private partnership, was launched in 1999.

Charting the course

The early proposals for genome-wide studies were audacious, because the number of SNPs known at the time was small, and the approaches to their discovery and genotyping were cumbersome. In 1998, Wang and colleagues performed an important feasibility study, discovering some 3,000 SNPs and developing an array-based genotyping approach that could assay hundreds of SNPs in parallel[23]. The SNP consortium and the HapMap project would eventually bridge the gap between this early survey and the much larger number of SNPs required.

How many SNPs are needed? The number of SNPs that are required for LD mapping obviously depends on the genomic extent of LD because genotyped SNPs must be spaced sufficiently densely to be in LD with most of the (potentially disease-associated) variants that are not genotyped. At the time the proposals for GWAS were made, few empirical estimates of the extent of LD were available, and these varied wildly from observations of LD over hundreds of kb to the breakdown of LD at very short distances. This range of observations translated into an uncertainty of up to three orders of magnitude in the required number of SNPs — from thousands to millions. Even several years later, the number of SNPs required for GWAS was said to be in the range of 30,000–1,000,000, based on a survey of empirical studies[24]. Starting in 1997, I attempted to reduce this uncertainty by using simple population-genetics models to calculate the likely extent of LD. A highly realistic model could not be constructed at the time because of a lack of detailed information regarding both the demographic history of different populations and the variation in recombination rate at short distances. Instead, the aim was to obtain a reasonable estimate. In the model that was designed to approximate the global human population, moderate levels of LD were confined to regions of approximately 6 kb, thus leading to the prediction that some 500,000 SNPs would be required for GWAS, even if relatively low LD levels between mapped SNPs and functional variants were deemed acceptable[25]. The predicted number of SNPs was considerably larger than the goals of SNP discovery projects at the time[26], and led to an increase in the targeted number. Indeed, less than 2 years later, a map of 1 million SNPs was reported[24]. Given the simplified nature of the model that was used to calculate the estimate of 500,000 SNPs, this number has held up remarkably well — most of the recent successes of GWAS came when approximately this number of SNPs could be genotyped within individual studies, and the current generation of commercial SNP-typing products deploys some 500,000–1,000,000 SNPs.

When the prediction is viewed from the vantage point of the extensive empirical data available today (for example, REF. 27),

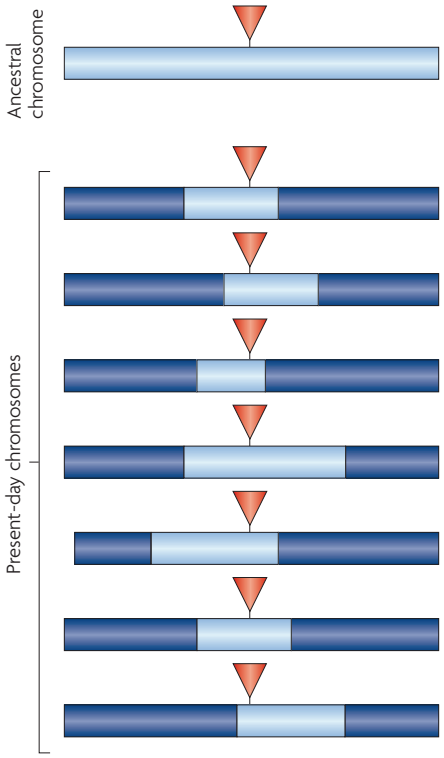

Figure 1 | **Linkage disequilibrium around an ancestral mutation.** The mutation is indicated by a red triangle. Chromosomal stretches that are derived from the common ancestor of all mutant chromosomes are shown in light blue, whereas new stretches introduced by recombination are shown in dark blue. Markers that are physically close (that is, within the light-blue regions of present-day chromosomes) tend to remain associated with the ancestral mutation, even as recombination whittles down the region of association over time.

Box 1 | **A brief history of genetic markers**

Human genetic mapping was initially based on restriction fragment length polymorphisms (RFLPs)[6,9,49,50] — fragment length variants generated through the presence and/or absence of restriction enzyme recognition sites[6,11,53,54]. RFLPs resulted from various sequence changes including base substitutions, insertions and deletions, and were laboriously assayed by Southern blots. Southern blots were superseded by PCR-based assays for microsatellite markers (also known as short tandem repeats or simple sequence length repeats)[10,51]. Microsatellites are di-, tri- or tetranucleotide repeat sequences that are composed of many tandem repeats. Many alleles are generally associated with each microsatellite within most populations, hence their use as markers for carrying out family-based linkage analysis[12,55]. More recently, SNPs have become the markers of choice; their lower polymorphism is offset by their abundance and ease of genotyping[23,56], and their low mutation rates make them especially suitable for linkage disequilibrium mapping.

it is clear that the actual average extent of LD is greater than in the model. This is especially true in non-African populations, most likely owing to a combination of demographic factors and a clustering of recombination events at hot spots[28] (both of these effects were anticipated when the prediction was made[25]). However, the calculation of the number of SNPs assumed both a relatively low acceptable level of LD and coverage of each region of LD with a single SNP. In practice, GWAS have set a higher standard for required LD, and multiple SNPs are used to 'tag' each region of high LD. These factors combined lead to the most current empirical estimates of approximately 500,000 SNPs for non-African and 1,000,000 SNPs for African populations to ensure adequate coverage of the genome in GWAS, even when high LD levels between mapped SNPs and functional variants are required[29].

How should the SNPs be chosen? Initially, the discussion focused on simply assembling a dense collection of SNPs. However, both theoretical considerations and early empirical studies suggested that the physical extent and the local patterns of LD were likely to vary across the genome and among populations. In commenting on one empirical study[30] in 1999, I proposed that LD among a dense collection of SNPs be measured empirically across the genome and in different populations in order to identify the most efficient SNP panels for association studies (FIG. 3); such panels would vary in their density by region of the genome and by population[31]. The result of such empirical studies would constitute an LD map of the human genome[31]. As SNP discovery efforts continued, empirical data confirmed both the need for hundreds of thousands of SNPs and the fact that these SNPs could not be chosen at random or by uniform spacing across the genome[5,32–35]. Rather, a million or more SNPs would need to be genotyped in substantial numbers of individuals from multiple populations in order to select sets of several hundred thousand (with the precise number depending on the population) that would efficiently capture untyped common variants[5,32–34]. These observations gave rise to the HapMap project and a parallel effort by Perlegen Sciences, which eventually joined forces to produce the SNP panels that are being used today[29,36,37]. These projects also drove the development of rapid and cost-effective genotyping technologies, setting the stage for GWAS. These recent developments are well chronicled elsewhere (for example, REF. 38).

The road ahead

In the past two decades, GWAS have progressed from visionary proposals, made when neither the sequence of the human genome nor many variations in this sequence were known, to routine practice of screening 500,000–1,000,000 SNPs in thousands of individuals. The recently reported phase 2 of the HapMap[29] now includes 3 million SNPs, estimated to cover one-quarter to one-third of all human SNPs with frequencies above 5%. Where do we go from here?

The recent crop of discoveries from GWAS is a major advance in our understanding of the genetic basis of common diseases, as well as normal human variation[39,40]. Nevertheless, the associated loci that have been identified usually have small individual effects on phenotype, and even collectively tend to explain only a small fraction of the heritable component[3]. For some diseases studied, no significant loci have been identified[3,41]. This failure to detect loci that explain the bulk of the heritable components of the phenotypes studied could be attributable to several factors. First, because the detected loci have small effects, the power to detect them is low, and more such loci remain to be discovered as sample sizes increase. Second, association studies can only detect the effects that are due to relatively common alleles. Rare alleles remain to be discovered — both at the loci that are identified by GWAS because they also have common alleles with phenotypic effects, and at other loci that do not have such common alleles. The former can be found by focused resequencing studies of the loci identified by GWAS; finding the latter might require resequencing of other genes in the relevant pathways, of the exons of all genes[42–44] or of the entire genome. Third, we might be missing the effects of structural variation, of other less well-studied types of genome alterations[45], and of interactions among variants and between genetic and environmental factors.

It is only a matter of time before all SNPs with appreciable frequencies in the human population have been discovered.

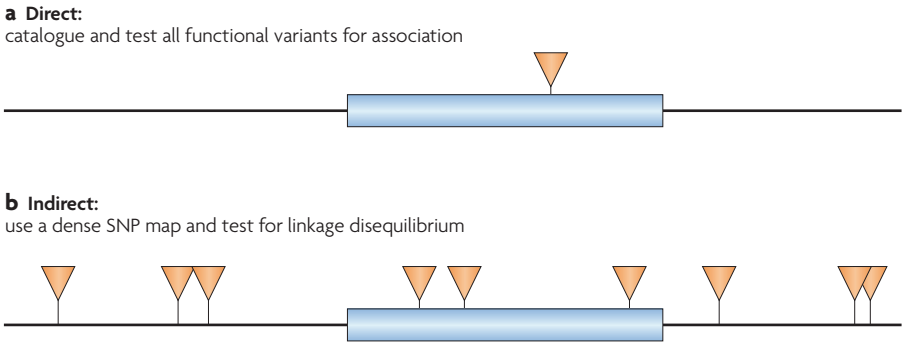

Figure 2 | **Alternative designs for genome-wide association studies. a** | Direct approach of testing a catalogue of all common functional variants in the genome. **b** | Indirect approach of testing a dense map of SNPs and relying on linkage disequilibrium to detect associations that are due to untested functional variants.

Indeed, efforts to discover and genotype additional SNPs in larger and more diverse population samples are underway. Assuming that the genotyping technologies can keep up the pace, the indirect association studies relying on LD will be replaced by direct association studies that assay all relatively common SNPs (perhaps the estimated 11 million SNPs with minor allele frequencies exceeding 1% in the population[46]), although it will probably still be worthwhile to exclude wholly redundant SNPs. Thus, LD and haplotype maps are merely useful but temporary shortcuts. An interesting finding of phase 2 of the HapMap is that for approximately 1% of all SNPs (tens of thousands), the basic assumption of indirect association mapping breaks down — these SNPs are not in LD with any others, often owing to their location in hot spots of recombination, and are thus 'untaggable' and must be assayed directly for phenotypic associations[29]. Tailored approaches that are under development will cover structural variants[47]. Studies based on the resequencing of individual genomes (rather than genotyping of known variants) will be needed to begin to comprehensively address the role of rare variants and *de novo* mutations, and will eventually replace genotyping studies altogether, although this is likely to take some time. It is worth noting that resequencing studies of rare variants have to rely on the recognition of many different variants, each of which alters the function of the same gene or pathway in different individuals. Whereas recognition of likely functional variants in coding regions is straightforward, detecting functional changes in noncoding DNA poses a major challenge. This is because regulatory sequences can be located far from the coding region and are often difficult to identify, and we do not have a ready connection between nucleotide differences and function for these sequences.

Looking further ahead, we can already envision the day when the genome sequences of a significant fraction of the population are known, at least in the developed world. Assuming that the relevant logistical and ethical issues can be solved, what will we learn from combining this unprecedented scope of genetic information with medical records and other phenotypic data? We are just beginning to get the first glimpses of the real underlying genetic complexity of phenotypic variation. Complexity is present on multiple levels, and might be fruitfully thought of as 'fractal'. First, many loci are involved; we do not yet know how many but the number could be in the hundreds for many traits. Second, individual loci can often represent variation in multiple linked genes, as has been found in model organisms (for example, REF. 48). Third, each gene is likely to contain multiple functional variants, including both 'superalleles' of linked alterations on one haplotype and allelic series with a range of allele frequencies and effect sizes. Non-additive interactions can be present at all levels. GWAS detect effects at the locus level, and an important challenge for future studies is identification of the genes, the functional variants and the functional mechanisms underlying phenotypic associations. Currently, such studies require painstaking, low-throughput experiments in cell lines and animal models.

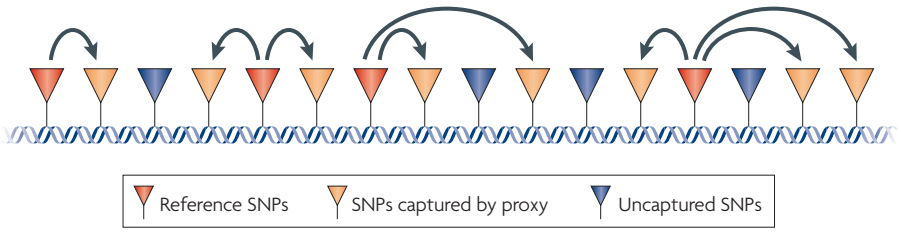

Figure 3 | **Schematic of a genomic region to be tested for association with a phenotype.** The four reference SNPs in the mapping panel are indicated by red triangles; these are genotyped directly. The eight SNPs indicated by yellow triangles are captured through linkage disequilibrium (by proxy) with the reference SNPs denoted by arrows. The four SNPs indicated by blue triangles are neither genotyped nor in linkage disequilibrium with the reference SNPs; phenotypic association that is due to one of these would be missed.

It is possible that some genetic contributions to human phenotypic variation might be too subtle to unravel, even when our surveys of the genome become truly comprehensive and the sample sizes approach that of the human population. Aside from the question of how much of the population variation we will ultimately be able to explain, we also have to ask how we can piece together individual risk from many small genetic contributions. Will we ultimately be able to classify individuals into meaningful groups with regard to risk of specific common diseases or response to drugs, as envisioned in personalized medicine? Doubtless this will be (or already is) true in some cases, but it is currently unknown how general such classifications are. We might need to replace some current phenotypic and disease classifications with ones that better correspond to the underlying genetic causes, perhaps by developing methods to iteratively refine phenotypic categories by combining genotypic and phenotypic information. Careful and detailed measures of phenotypes and environmental exposures will also have an important role. Clearly, we have a lot of work to do before an individual genome sequence is more phenotypically informative than it is today[49,50]. In the meantime, great care is required in offering genome-based information to individuals[51,52].

Concluding remarks

What is the best future direction for human genetics? There are essentially three avenues to pursue: much larger samples; better assays of genome variation that can capture both common alterations that are not in LD with SNP panels and rare variants; and more detailed phenotyping. Undoubtedly, each of these approaches has a role, and we do not yet have all the information needed to decide which will prove most fruitful. Therefore, it is a high priority to apply a full battery of approaches to several model diseases and phenotypes in order to empirically determine the range of outcomes, just as the Wellcome Trust Case Control Consortium study of seven diseases provided an empirical guide for GWAS[41]. In my opinion, the most pressing question is the contribution of rare variants, both in the genes that harbour common risk variants and in those that do not. This question is also the most difficult to address comprehensively with today's technologies, but it seems imperative that we prioritize studies to begin to get the answers.

Leonid Kruglyak is at the Lewis-Sigler Institute for Integrative Genomics and the Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA.

e-mail: leonid@genomics.princeton.edu

doi:10.1038/nrg2316
Published online 19 February 2008

1. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* **39**, 813–815 (2007).
2. Bowcock, A. M. Genomics: guilt by association. *Nature* **447**, 645–646 (2007).
3. Gibson, G. & Goldstein, D. B. Human genetics: the hidden text of genome-wide associations. *Curr. Biol.* **17**, R929–R932 (2007).
4. Topol, E. J., Murray, S. S. & Frazer, K. A. The genomics gold rush. *JAMA* **298**, 218–221 (2007).
5. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
6. Botstein, D., White, D. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
7. Solomon, E. & Bodmer, W. F. Evolution of sickle variant gene. *Lancet* **1**, 923 (1979).
8. Lander, E. S. & Botstein, D. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 49–62 (1986).
9. Kan, Y. W. & Dozy, A. M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc. Natl Acad. Sci. USA* **75**, 5631–5635 (1978).
10. Bodmer, W. F. Human genetics: the molecular challenge. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 1–13 (1986).
11. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
12. Weissenbach, J. *et al.* A second-generation linkage map of the human genome. *Nature* **359**, 794–801 (1992).
13. Kerem, B. -S. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
14. Houwen, R. H. J. *et al.* Genome scanning by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8**, 380–386 (1994).
15. Puffenberger, E. G. *et al.* Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.* **3**, 1217–1225 (1994).
16. Risch, N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* **46**, 242–253 (1990).
17. Risch, N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* **46**, 229–241 (1990).
18. Risch, N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* **46**, 222–228 (1990).
19. Kruglyak, L. & Lander, E. S. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**, 439–454 (1995).
20. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
21. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
22. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
23. Wang, D. G. *et al.* Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
24. The International SNP Map Working Group. A map of human genome sequence variation containing 1 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
25. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
26. Collins, F. S. *et al.* New goals for the US human genome project: 1998–2003. *Science* **282**, 682–689 (1998).
27. Pe'er, I. *et al.* Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am. J. Hum. Genet.* **78**, 588–603 (2006).
28. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
29. The International HapMap Consortium. A second generation human haplotype map of over 3 million SNPs. *Nature* **449**, 851–861 (2007).
30. Lonjou, C., Collins, A. & Morton, N. E. Allelic association between marker loci. *Proc. Natl Acad. Sci. USA* **96**, 1621–1626 (1999).
31. Kruglyak, L. Genetic isolates: separate but equal? *Proc. Natl Acad. Sci. USA* **96**, 1170–1172 (1999).
32. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
33. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
34. Reich, D. E., Gabriel, S. B. & Altshuler, D. Quality and completeness of SNP databases. *Nature Genet.* **33**, 457–458 (2003).
35. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
36. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
37. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
38. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
39. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genet.* (2007).
40. Weedon, M. N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nature Genet.* **39**, 1245–1250 (2007).
41. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
42. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
43. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* (2007).
44. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
45. Legendre, M., Pochet, N., Pak, T. & Verstrepen, K. J. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* **17**, 1787–1796 (2007).
46. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
47. Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3**, 1787–1799 (2007).
48. Sinha, H., Nicholson, B. P., Steinmetz, L. M. & McCusker, J. H. Complex genetic interactions in a quantitative trait locus. *PLoS Genet.* **2**, e13 (2006).
49. Brenner, S. E. Common sense for our genomes. *Nature* **449**, 783–784 (2007).
50. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
51. Anonymous. Risky business. *Nature Genet.* **39**, 1415 (2007).
52. McGuire, A. L., Cho, M. K., McGuire, S. E. & Caulfield, T. Medicine. The future of personal genomics. *Science* **317**, 1687 (2007).
53. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
54. Wyman, A. R. & White, R. W. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77**, 6754–6758 (1980).
55. Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
56. Kruglyak, L. The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* **17**, 21–24 (1997).

Acknowledgements

I thank many colleagues over the years, the participants at the 2007 Banbury Center Meeting "From Statistics To Genes: Figuring Out the Molecular Basis of Complex Traits", D. Altshuler for discussions, and D. Botstein, A. Chakravarti, B. Coller, D. Goldstein and L. Rosenberg for comments on the manuscript. I regret that space constraints prevented me from citing other important work in the field. Supported by a MERIT award from the National Institutes of Health (R37 MH059520) and a James S. McDonnell Centennial Fellowship in Human Genetics.

DATABASES

OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
benign recurrent intrahepatic cholestasis (BRIC) | Hirschsprung disease

FURTHER INFORMATION

Leonid Kruglyak's homepage: http://www.molbio2.princeton.edu/index.php?option=content&task=view&id=217
Perlegen Sciences: http://www.perlegen.com
The International HapMap project: http://www.hapmap.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF