Biochemical and Biophysical Research Communications 396 (2010) 143–146

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Genetics of complex disorders

Juha Kere

Karolinska Institutet, Department of Biosciences and Nutrition, 14157 Huddinge, Sweden
Folkhälsan Institutet of Genetics, Helsinki, and Department of Medical Genetics, University of Helsinki, 00014 Helsinki, Finland

ARTICLE INFO

Article history:
Received 31 March 2010

Keywords:
Gene mapping
Linkage disequilibrium
Genome-wide association study

ABSTRACT

The success stories of identifying genes in Mendelian disorders have stimulated research that aims at identifying the genetic determinants in complex disorders, in which both genetics, environment and chance affect the pathogenetic processes. This review summarizes the brief history and lessons learned from genetic analysis of complex disorders and outlines some landscapes ahead for medical research.

1. A short history of rapidly changing analytic methods

The history of modern human genetics research is much the history of the rapidly changing methodologies. The vast size of the human genome appeared at first an impossible task to work with, and certainly to know in full detail. But with the advent of new analytical approaches roughly every 10 years, we have now reached the point where individual genome sequencing is within the reach for many laboratories.

The brief history and some of the key papers to read related to each stage are listed in Table 1. Modern human molecular genetics research owes much to the innovative method that bears the name of its inventor: Southern blotting (and the analogous gel-blotting techniques named with a tongue in the cheek, Northern and Western blotting). More exactly, Southern blotting applied to the analysis of restriction fragments obtained after digestion with a growing list of various bacterial restriction enzymes allowed the analysis of genome position-specific polymorphic markers [1]. The idea that these markers, restriction fragment length polymorphisms (or RFLPs) could be used as anonymous addresses for the various positions in the human genome led to systematic efforts to clone probes and screen them for their polymorphic nature [2]. Ultimately, hundreds of such cloned probes combined with the study of large families were used to derive the first continuous genetic maps of the human genome [3]. Such maps then made it possible to attempt the mapping of any Mendelian trait to a named position in the genome, with first success stories setting the stage of the first wave of disease gene identifications [10].

Southern blotting combined with the use of radioactive probes was not optimal for large-scale projects. The invention of the polymerase chain reaction (PCR) changed molecular genetic analyses rapidly, and a new class of genetic markers was discovered. When RFLPs were caused either by single nucleotide changes altering restriction enzyme recognition sites or larger-scale variations in tandemly arranged DNA repeat units, the new class of microsatellite markers were based on length variation in short sequence repeats [4]. This class of markers, often involving poly-CA dinucleotide repeats, or various tri- or tetranucleotide repeats, could be analyzed more rapidly and easily, and moreover, there appeared to be many more of them in the genome. The second-generation genetic linkage maps of the human genome were then based entirely on PCR-amplified microsatellite markers [5].

The analysis of microsatellite markers still required the separation of differently-sized DNA fragments. Another line of technology was being developed based on the high selectivity of DNA hybridization that could discern sequence differences as small as a single nucleotide. The rather common and simple assay based on the application of short DNA probes on filter papers as dot-blots was refined and miniaturized to make microarrays of DNA probes on glass slides. The first applications of the microarray technology were directed toward measuring levels of different transcripts in RNA samples [11]. When the common variation in DNA sequences between individuals denoted single-nucleotide polymorphisms (SNPs) started to be appreciated, it became possible to make microarrays assaying at first thousands and later hundreds of thousands of SNPs all over the genome that were discovered in the meantime by genome sequencing and other means.

In parallel with these developments, DNA sequencing technologies took rapid leaps. The dideoxynucleotide-based sequencing method invented by Fred Sanger [12] was taken from manually-read radioactively labelled gels to machine-read fluorescently labelled gels and ultimately to matrix-filled capillaries. Massively parallel methods based on immobilized DNA molecules on the surface of microparticles then evolved in various forms to the new

E-mail address: juha.kere@ki.se

0006-291X/$ - see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.bbrc.2010.04.013

Table 1
Selected key events in the mapping of disease genes.

Development	Importance	References
DNA blotting	Easy analysis of DNA variation	[1]
Concept of RFLP maps	Vision of human gene mapping	[2]
Continuous linkage maps of the human genome	Genome-wide linkage mapping of human genes	[3]
Microsatellite maps	Mapping of many human disease genes	[4,5]
Extended linkage disequilibrium in the genome	Vision of linkage disequilibrium mapping in complex disorders	[6,7]
The HapMap project	Detailed analysis of linkage disequilibrium in the human genome	[8]
Genome-wide association studies in complex disorders	Hundreds of loci identified	[9]

generation DNA sequencing technologies. Again, one example of a yet newer technology is based on assaying the progressing synthesis of a DNA strand of individual DNA polymerase molecules on nanofabricated devices [13].

2. Candidate gene studies

All these methodologies were first applied to the study of monogenic or Mendelian disorders. Even though complex disorders seemed a difficult target for study, the early attempts to disentangle the genetic determinants were based on candidate gene approaches. The rationale was to use genetic association to ask whether certain polymorphisms in genes that were thought to be functionally relevant in a given disease would show different allele frequencies in patients as compared to controls. A large number of studies were published, many based on patient and control collections as small as only tens of samples, or at best several hundreds, and numerous weak associations were reported in the literature. Some of the identified genes have withstood the verification by better-powered genetic association studies, but most have never been replicated in another study.

3. Chromosomal markers as pointers to genes in complex disorders

In complex as well as in monogenic disorders, one way to identify possible disease-related genes has been to study specific rare chromosome abnormalities in patients with a monogenic or complex disorder. In monogenic diseases, there are numerous success stories, such as for example the mapping and identification of the Duchenne muscular dystrophy gene or the anhidrotic ectodermal dysplasia gene [14,15]. In complex disorders, however, some of the identified genes have likewise turned out to be relevant even by genetic association tests; the first dyslexia gene DYX1C1 might serve as an example [16].

4. Genetic linkage studies

Genetic linkage studies in monogenic disorders are based on the simple model of Mendelian inheritance, and linkage analysis algorithms based on this model are very powerful for showing genetic linkage. In complex disorders, however, the situation is much less straightforward as a given genotype would not at all fully correlate with the presence of the disease phenotype. In genetic terms, both phenocopies (the occurrence of the same phenotype in genetically completely different individuals), reduced penetrance (the occurrence of the phenotype in only a fraction of those possessing a given genotype) and variable expressivity (the clinical heterogeneity of a given condition in the presence of the same genotype in different individuals) complicate matters. Therefore, new computational approaches were needed to complement those of model-based monogenic linkage studies.

Many successful approaches in complex disorders were based on affected-only strategies. The idea was to identify genomic regions that more often than by chance were shared between individuals affected with a complex disorder [17]. No information content was given to healthy individuals. Even though phenocopies might hamper the identification of shared genomic regions, large enough sample sets should ultimately implicate specifically disease-linked positions with acceptable levels of statistical evidence.

The 1990s were the golden period of genome-wide linkage studies in complex disorders. Accumulated studies, however, started little by little to cast doubt on the replicability of the mapping results. Typically, no two studies of the same disorder reported consistently the same loci, but instead all chromosomes were littered with positive but nonreplicated linkage hits.

Only quite recently has it become apparent what was the problem with these approaches. We now know that many true susceptibility genes in complex disorders have effect sizes so small that genetic linkage, in order to be powerful enough should have been based two or three orders of magnitude larger sample sets. The early attempts to map genes in complex disorders were not wrong in principle, but the studies were simply underpowered to detect the weak genetic effects underlying many complex disorders.

Nevertheless, strong single gene effects in some complex disorders have been abundantly verified already by genetic linkage and later confirmed by genetic association. An illustrative example is psoriasis, where the overwhelming genetic effect of the HLA-C gene or SNPs nearby was detected in virtually all genetic studies [18].

5. The paradigm of positional cloning in complex disorders

In monogenic disorders, the genetic mapping of the causative locus was followed up by narrowing down the DNA segment including the causative gene by genetic association. The idea was to take advantage of founder effects, especially in recessive disorders where a mutated gene might be inherited for generations in a population until it happened to find its way homozygous in an individual, then causing disease. The same principle was applicable to the thinking in complex disorders [19]. A susceptibility allele with reduced penetrance might travel for generations in families and populations with little if any selection, as many complex disorders only manifest after the reproductive age window.

Some success stories exist with this approach. One example is the genetic mapping and positional cloning of NPSR1 as a susceptibility gene for asthma and related disorders [20]. In these studies, we took advantage of a rural isolated subpopulation in Finland hoping to reduce the genetic complexity of an otherwise heterogeneous complex disorder. By a lucky incidence, a particular asthma risk associated variant appeared to be more common there than in many other populations, possibly helping to reach barely sufficient power to map the gene by linkage and association. Later association studies, including well-powered genome-wide association studies have confirmed the disease association of NPSR1 to asthma.

6. The block structure of human chromosomes and the HapMap project

Millions of SNPs across the genome are not particularly helpful to map and identify disease-associated genes in complex disorders,

if they all need to be assayed in thousands of individuals to detect association signals. As late as the end of the 1990s, the human genome was thought to have undergone a large number of random recombination events that reduced the length of possibly associated fragments to only a few thousand base pairs. Given the size of the genome, it was predicted that at least 500,000 SNPs would be needed to perform a genome-wide association study with enough coverage of the genome [21].

A radical change in thinking was imposed by the emerging observation that there were much longer segments of strong genetic association between SNPs than anticipated by the random recombination model [6,7]. The observation from a limited number of short segments in various positions of the genome was then expanded to a genome-wide hypothesis of so-called haplotype blocks. Their origin was explained to be in the short history of the human population in its rapidly growing expansive phase, not allowing the magnitude of recombination to have taken place that was earlier thought. Whether the segmented recombination events were caused by hotspots in the DNA sequence or simple by historical coincidences became a matter of debate, but the observation remained that especially SNPs with both alleles relatively common, inferred to be old, usually show long distances of genetic association, or linkage disequilibrium, along chromosomes.

An immediate consequence of this new image of the fine structure of the human genome sequence was that much fewer SNPs would be needed to assess genetic association to old susceptibility alleles that would be sufficiently common in the populations, thus old, and thus in strong linkage disequilibrium with many nearby SNPs.

An international project named the HapMap project was initiated to systematically identify a large number of SNPs, genotype them in sets of samples collected from major population groups from three different continents, and to calculate the levels of genetic association between all pairs of nearby markers [8]. The results of this project guided the design of microarrays that subsequently became major tools for genome-wide association studies.

7. Genome-wide association studies

The new paradigm for identifying genes involved in complex disorders was based on the idea that if a SNP allele was either itself causally involved in a disease, or in sufficiently tight linkage disequilibrium with a causative variant, its altered frequency in a large number of patients compared to controls would flag the susceptibility gene. It is important to note that genes with multiple rare mutations would not fulfill any of these two prerequisites, and would thus not be detected by genome-wide association studies.

After a few earlier success stories, a landmark study marked the avalanche of genome-wide association studies in a large variety of complex disorders [9]. Hundreds of susceptibility genes have now become implicated with p values reaching astronomical levels, and even larger numbers of loci remained just below significance levels across the genome. New principles of calling a genetic association genome-wide significant were agreed on, and it became obvious that most of the earlier candidate gene-based studies most likely were false positives because of insufficient power caused by small sample size. Increasingly large studies based on international consortia were launched and executed, and the number of loci increased rapidly. At this writing, over 500 genome-wide association studies have been published that have used at least 100,000 SNPs to screen the genome and have resulted in association p values better than 10^{-5} (http://www.genome.gov/26525384).

8. The dark matter of inheritance

A new and completely unexpected problem started to unveil. The larger the genome-wide studies became, the weaker genetic effects they could record with significant p values, but at the same time it became apparent that taken together, the implicated variants explained only a small fraction of the genetic component in the diseases under study [22]. There appeared to be a genetic component that was not mappable by genome-wide association studies, and the term "dark matter of inheritance" was coined, analogous to the invisible matter in the universe.

Good examples are provided by studies that have identified loci associated with human height. Adult height is a highly heritable trait, with heritability estimates mostly well over 0.8. It is also relatively easy to measure accurately, allowing the combination of large population samples from different sources. With increasingly large studies, six including more than 10,000 individuals, several tens of genes have been implicated, each contributing with only a small effect. In one illustrative study reporting 20 loci, they explained combined about 3% of height variation [23]. Furthermore, the 20 loci did not show any evidence of gene $\times$ gene interaction, or epistatic effects, but the length of individuals was linearly proportional to the number of "tall" alleles. Similar results have been reported for many quantitative traits, such as body mass index or serum triglyceride and high density lipoprotein levels.

What might then explain the heritability tat remains invisible to genome-wide association studies? There might be still many more small effect-size variants that have not yet been discovered. Several possible technical explanations have been proposed, including rare variants not covered by current genotyping microarrays containing mostly variants with population frequencies <5%, or structural DNA variants that are not identified using microarrays. Statistical explanations include the failure to detect gene $\times$ gene interactions and the omission of environmental data of possible importance for gene $\times$ environment interactions. One biological explanation should consider the possibility of epigenetic effects that might be modified by environmental effects but modify chemically DNA, e.g., by methylation of important regulatory regions.

9. Landscapes ahead

It is evident that the rapid development of DNA sequencing technologies and the continuously improving microarrays contribute to an extremely expansive phase of biological research. The biological revolution not merely continues; it accelerates in an explosive fashion.

It is anticipated that DNA sequencing will rapidly become the major method of genetic analysis. For the past 2 years, the raw DNA sequence output of the next-generation sequencing instruments has been growing faster than Moore's law predicted for computers. At least one company using a combinatorial DNA sequencing technology already today promises to deliver an assembled full genome sequence for less than 5000$, and the price tag of 1000$ is likely to be reached very soon multiple technologies. A newly commercially launched technology that is based on real-time recording of the synthetic steps of individual DNA polymerase molecules on nanofabricated devices delivers now single molecule sequence reads up to 6 kb in length. The next generation devices are likely to improve on all current numbers for performance and cost. It is difficult to think of a field of genomic or genetic analysis that could not use advanced DNA sequencing with advantage over current methods.

The 1000 genomes project will give us a much more complete picture of rare variation than seen before. Very soon, the 1000

completely sequenced genomes will be extended to tens of thousands of individually sequenced genomes, and the results are likely to tell us that there is a vast amount of not only rare but individual variation as expected based on population genetic considerations [24]. This will complicate genetic analysis. Clustering of several private deleterious mutations in the same gene in a group of patients will suggest functional relevance, but statistical arguments or other conclusions cannot easily be made on the basis of single individuals with mutations in poorly characterized genes even if the mutations would be deleterious to function. We already have the first estimate of the overall mutation rate in the human genome that is based on the direct comparison of parents' and child's full genomic sequences, suggesting that on the average 30 new mutations are passed by each parent to the zygote [25]. Out of the 60 or so new mutations, which one is the right one to explain a rare disorder?

Individual exomic sequencing of a handful of cases with rare monogenic disorders has already turned out to be successful in identifying causative genes, even in the absence of founder effects [26]. Can this analysis be extended to complex disorders? If so, how many samples are likely to be needed for exomic or complete sequencing? Will we be able to see clustering of private mutations in certain genes? How can we best combine transcriptome sequencing to detect pathophysiological changes in complex disorders? These questions are already under intensive investigation.

Further studies will also look at the role of epigenetic modifications on gene functions and genetic associations. Our unpublished results as well as those of others show that a methylated cytosine may alter transcription factor binding in comparison to an unmethylated cytosine in a functional promotor element [27]. This observation has consequences even for genetic association studies. If a methylated C allele behaves functionally like, e.g., a T allele, and different from an unmethylated C allele, a simple genetic analysis result will become diluted, because the methylation information of the C alleles is neglected. Thus, it may become necessary to carefully scrutinize CpG dinucleotides in disease-associated genes for possible methylation and functional effects, and ultimately correct genetic effect estimates appropriately.

In a historically short period of time, genetic analysis has evolved from merely collecting rare victories by time-consuming manual analyses to automated analyses driven by technologies involving highly parallelized and miniaturized assays. Data production as measured, e.g., by stored DNA sequences has reached an accelerating expansive phase (http://www.ncbi.nlm.nih.gov/Genbank). This development imposes great demands on our intellectual capacity to handle the information. Biology and medicine have turned into disciplines that are today characterized more by our capacity to ask meaningful questions and answer them by processing the available data, rather than our perseverance to perform months or years of difficult experimentation. The consequences for medical research, our understanding of disease processes, and opportunities to design new treatments will be beyond the wildest fantasies as Karolinska Institutet completes its first 200 years of medical research.

References

[1] E.M. Southern, Detection of specific sequences among DNA fragments separated by gel electrophoresis, J. Mol. Biol. 98 (1975) 503–517.
[2] D. Botstein, R.L. White, M. Skolnick, R.W. Davis, Construction of a genetic linkage map in man using restriction length polymorphisms, Am. J. Hum. Genet. 32 (1980) 314–331.
[3] H. Donis-Keller, P. Green, C. Helms, S. Cartinhour, B. Weiffenbach, K. Stephens, T.P. Keith, D.W. Bowden, D.R. Smith, E.S. Lander, et al., A genetic linkage map of the human genome, Cell 51 (1987) 319–337.
[4] J.L. Weber, P.E. May, Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction, Am. J. Hum. Genet. 44 (1989) 388–396.
[5] NIH/CEPH Collaborative Mapping Group, A comprehensive genetic linkage map of the human genome, Science 258 (1992) 148–162.
[6] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, E.S. Lander, High-resolution haplotype structure in the human genome, Nat. Genet. 29 (2001) 229–232.
[7] D.E. Reich, M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, E.S. Lander, Linkage disequilibrium in the human genome, Nature 411 (2001) 199–204.
[8] International HapMap Consortium, The International HapMap Project, Nature 426 (2003) 789–796.
[9] Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls, Nature 447 (2007) 661–678.
[10] J.F. Gusella, N.S. Wexler, P.M. Conneally, S.L. Naylor, M.A. Anderson, R.E. Tanzi, P.C. Watkins, K. Ottina, M.R. Wallace, A.Y. Sakaguchi, A.B. Young, I. Shoulson, E. Bonilla, J.B. Martin, A polymorphic DNA marker genetically linked to Huntington's disease, Nature 306 (1983) 234–238.
[11] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.
[12] F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.
[13] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al., Real-time DNA sequencing from single polymerase molecules, Science 323 (2009) 133–138.
[14] A.P. Monaco, R.L. Neve, C. Colletti-Feener, C.J. Bertelson, D.M. Kurnit, L.M. Kunkel, Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene, Nature 323 (1986) 646–650.
[15] J. Kere, A.K. Srivastava, O. Montonen, J. Zonana, N. Thomas, B. Ferguson, F. Munoz, D. Morgan, A. Clarke, P. Baybayan, et al., X-linked anhidrotic (hypohidrotic) ectodermal dysplasia is caused by mutation in a novel transmembrane protein, Nat. Genet. 13 (1996) 409–416.
[16] M. Taipale, N. Kaminen, J. Nopola-Hemmi, T. Haltia, B. Myllyluoma, H. Lyytinen, K. Muller, M. Kaaranen, P.J. Lindsberg, K. Hannula-Jouppi, et al., A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain, Proc. Natl. Acad. Sci. USA 100 (2003) 11553–11558.
[17] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, E.S. Lander, Parametric and nonparametric linkage analysis: a unified multipoint approach, Am. J. Hum. Genet. 58 (1996) 1347–1363.
[18] International Psoriasis Genetics Consortium, The International Psoriasis Genetics Study: assessing linkage to 14 candidate susceptibility loci in a cohort of 942 affected sib pairs, Am. J. Hum. Genet. 73 (2003) 430–437.
[19] E.S. Lander, N.J. Schork, Genetic dissection of complex traits, Science 265 (1994) 2037–2048.
[20] T. Laitinen, A. Polvi, P. Rydman, J. Vendelin, V. Pulkkinen, P. Salmikangas, S. Mäkelä, M. Rehn, A. Pirskanen, A. Rautanen, et al., Characterization of a common susceptibility locus for asthma-related traits, Science 304 (2004) 300–304.
[21] L. Kruglyak, Prospects for whole-genome linkage disequilibrium mapping of common disease genes, Nat. Genet. 22 (1999) 139–144.
[22] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, et al., Finding the missing heritability of complex diseases, Nature 461 (2009) 747–753.
[23] M.N. Weedon, H. Lango, C.M. Lindgren, C. Wallace, D.M. Evans, M. Mangino, R.M. Freathy, J.R. Perry, S. Stevens, A.S. Hall, et al., Genome-wide association analysis identifies 20 loci that influence adult height, Nat. Genet. 40 (2008) 575–583.
[24] M.E. Zwick, D.J. Cutler, A. Chakravarti, Patterns of genetic variation in Mendelian and complex traits, Annu. Rev. Genom. Hum. Genet. 1 (2000) 387–407.
[25] J.C. Roach, G. Glusman, A.F. Smit, C.D. Huff, R. Hubley, P.T. Shannon, L. Rowen, K.P. Pant, N. Goodman, M. Bamshad et al., Analysis of genetic inheritance in a family quartet by whole-genome sequencing, Science (Epub ahead of print).
[26] S.B. Ng, E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E.E. Eichler, et al., Targeted capture and massively parallel sequencing of 12 human exomes, Nature 461 (2009) 272–276.
[27] M.R. Campanero, M.I. Armstrong, E.K. Flemington, CpG methylation as a mechanism for the regulation of E2F activity, Proc. Natl. Acad. Sci. USA 97 (2000) 6481–6486.