Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 8315–8317, July 1999

Commentary

A paradigm for finding genes for a complex human trait: Polycystic ovary syndrome and follistatin

Kunle Odunsi and Kenneth K. Kidd*[†‡]

**Department of Gynecological Oncology, Roswell Park Cancer Institute, Buffalo, NY 14263; and [†]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520*

Positional cloning using genetic linkage has been very successful for many single-gene Mendelian traits. In contrast, complex disorders with unknown but largely genetic etiology (although not a simple Mendelian inheritance pattern) are just beginning to yield to a positional cloning approach. For many such disorders, very large numbers of affected sibling pairs appear to be required to have any hope of significant results on a genomewide survey for increased allele sharing. In this issue of the *Proceedings*, the study by Urbanek *et al.* (1) of polycystic ovary syndrome (PCOS) shows that a different approach is possible when there is a good understanding of the physiology and relevant metabolic pathways in the affected tissues, i.e., a survey of allele sharing at markers in or adjacent to candidate genes. In contrast to the flailing about with a few poorly supported candidate genes seen in psychiatric genetics, Urbanek *et al.* (1) examined all of the plausible candidate genes based on a good understanding of metabolism and physiology in ovarian tissues. They were able to use a relatively small sample to obtain quite convincing evidence that one of the genes, and possibly a second, are likely to have major effects in the etiology of PCOS.

The major mechanisms responsible for several complex human diseases are poorly understood. At least four major obstacles hinder the identification of the major defect in these conditions. The first is the lack of recognition that they are syndromes and not individual diseases; the second is the variety of possible etiologies responsible for disease; the third is the complex interplay of several physiologic systems that regulate the final clinical manifestation of disease; and the fourth is the interaction between environment and disease. For genetic diseases that are not associated with obvious structural rearrangements of chromosomes, the causative gene(s) can be localized by genetic linkage analysis in families segregating for the disease phenotype. Genetic linkage is the phenomenon whereby loci appear to be transmitted together rather than independently to offspring. For a complex disorder, analysis involves testing whether an allele at a marker locus is preferentially transmitted along with the disorder through a family. If there is statistically significant cotransmission of marker and disorder, the explanation is that some genetic variant in the chromosomal segment around the marker is contributing to the development of the disorder.

The paper by Urbanek *et al.* (1) describes genetic linkage analysis on an extremely important endocrinopathy, PCOS. The full-blown syndrome of hyperandrogenism (HA), chronic anovulation, and polycystic ovaries affects up to 5–10% of all premenopausal women (2); other features may include obesity, hirsuitism, and hyperinsulinism as well as increased risk for developing non-insulin-dependent diabetes mellitus (3), atherosclerosis, hypertension, dyslipidemia (4), coronary artery disease, and endometrial carcinoma. Since the initial description of the syndrome by Stein and Leventhal in 1935 (31), the diagnostic criteria have been remarkably modified so much that polycystic ovaries, the criterion that was originally the *sine qua non* of the syndrome, has become a "consistent finding" rather than an essential diagnostic criterion. There is not one homogenous group of PCOS patients, but rather a spectrum of patients sharing, for the most part, the same clinical features that have arisen by similar but probably diverse etiopathophysiologic processes (5). It is likely that hyperinsulinemic hyperandrogenism represents a significant subgroup, present in >50% of patients with hyperandrogenism of chronic anovulation, and this group can be obese or nonobese. Thus, the PCOS phenotype is complex, and genetic analysis will necessarily require an understanding of the possible physiologic mechanisms of the disease to search for candidate genes. Although the exact mechanism for the development of PCOS is not known, evidence indicates that alterations in the endocrine, paracrine, and autocrine control of ovarian folliculogenesis are involved.

Three major theories have been proposed to explain the cause of PCOS. First, the luteinizing hormone–theca interstitial cell (LH-TIC) theory suggests that the pathophysiologic mechanisms leading to abnormally elevated levels of LH underlie the phenomenon of PCOS. The theory suggests that high levels of circulating LH cause an increase in the growth of TIC in developing follicles, which leads to androgen overproduction and follicular atresia. The second theory, the follicle stimulating hormone–granulosa cell (FSH-GC) theory suggests that FSH leads to subnormal induction of cytochrome P450 aromatase in the granulosa cell, leading to elevated androgen levels. This may be due to insufficient bioactive FSH in the follicular microenvironment to induce P450 aromatase gene expression, dysfunctional FSH receptor signal transduction mechanism, or the presence of inhibitors [such as epidermal growth factor and insulin-like growth factor (IGF)-binding protein 3] that prevent the normal expression of P450 aromatase activity. The third theory relates to the growth factor–autocrine paracrine system. In PCOS, there is evidence of an altered IGF/insulin system, and these act as mediators of biologic responses of the selectogenic and atretogenic follicular hormones (6–8). The analysis by Urbanek *et al.* (1) covered all possible candidate genes that may be involved in the pathways of the major theories discussed above. These include steroid hormone metabolism and action, gonadotropic action, obesity and energy regulation, and insulin action.

The design of the study by Urbanek *et al.* (1) and the clear presentation of the results serve as a model for studies of this kind, especially when the disease phenotype is complex. HA was used in the paper to define a "homogenous" subgroup of PCOS, and this led to a greater power to detect the loci contributing to HA. If menstrual irregularity is an epistatic phenotype "built" on HA, there will be less power to detect such loci. In fact, the use of such an "intermediate" phenotype may prove crucial in elucidating the physiologic mechanism(s) by which an identified relevant genetic variant imparts an

The companion to this Commentary begins on page 8573.
[‡]To whom reprint requests should be addressed.

effect (9). Thus, in addition to the well understood metabolic pathways and the genes controlling those pathways, the authors were able to define a subgroup likely to be more etiologically homogenous. For other complex disorders with no dichotomous biochemical test, success in defining such subgroups is often a matter of luck and is sometimes the result of clinical expertise and insight. Here, knowledge and clinical expertise were operative.

Often, the characterization of gene(s) involved in complex diseases is a difficult task and may require substantial resources, large family data, highly polymorphic genetic markers that span the genome and specifically developed statistical approaches. For many complex diseases, traditional logarithm of odds (lod) score analysis is unlikely to be powerful, because it assumes the presence of a single, major disease locus (with a specific mode of inheritance) that accounts for the majority of genetic variance. Although this can be remedied by carrying out parametric lod score analyses under different disease models while allowing for heterogeneity (10, 11), model-free nonparametric methods are also possible. One such method is the "affected-sibling-pair" (ASP) method, which tests for increased marker similarity in affected sibling pairs. The method requires no assumptions about the mode of inheritance, but is most powerful when identity-by-descent of marker alleles can be determined explicitly and need not be statistically estimated. Genetic linkage analysis requires the use of polymorphic markers. Almost all human DNA polymorphisms are based on single-nucleotide substitutions or on variations in the number of tandem repeats. Currently, for linkage mapping of genetic disease genes and for construction of high-resolution linkage maps, short tandem repeat polymorphisms (STRPs), usually dinucleotide repeats (mostly CA repeats), but also tri- and tetranucleotide repeats, are the markers of choice because of their high degree of polymorphism (several alleles at each locus) and their abundance in the genome. As Urbanek *et al.* (1) discussed, it was often possible to unambiguously determine the parental genotype(s) for the STRPs even when the parents were not available, thereby allowing unambiguous classification of sibling pairs by number of alleles shared identical by descent (IBD).

The power to detect linkage by using the ASP method is a function of the contribution of the locus to the genetic variation of the trait. This can be measured as the risk ratio (λs), the risk to a sibling of an affected proband versus the population prevalence (12–14). λs is an overall risk ratio that summarizes the collective effect of all disease loci and for complex diseases. The question at the beginning of this type of study is: For a given sample size, what is the minimum λs for which we have a good chance of detecting linkage? Power estimates assume that the marker and disease-susceptibility locus are linked and that the marker is informative. The effects of recombination can be reduced by multimarker analysis (15–17) whereas the effects of noninformativeness can be reduced by using highly polymorphic markers. The closer the polymorphism is to the disease gene, the less likely is recombination between the two at meiosis, and therefore the two are more likely to be inherited together. High-density genetic maps that facilitate positional cloning projects have been generated by using STRPs (18, 19). For 28 of the 37 candidate genes analyzed by Urbanek *et al.* (1), there is at least one polymorphic marker within 1 centiMorgan (cM) of the candidate gene, and for the remainder, the markers are 1–4 cM from the candidate gene. This proximity improves the power of the study because recombination is likely to be minimal.

When candidate genes are not known, or in addition to the candidate gene approach, linkage studies can be performed by using anonymous markers. A series of markers that appropriately span the entire genome is used in an attempt to determine those areas in which a susceptibility gene may reside. The difficulty with whole-genome scans is the level of significance required to adequately compensate for the multiple independent and partially independent tests carried out. Although the required significance levels can be determined (17), the sample sizes can be prohibitive. Urbanek *et al.* (1) avoided this problem by focusing on a relatively smaller set of candidate genes.

In the paper by Urbanek *et al.* (1), the IBD levels observed for D5S623, the marker closest to follistatin ($<$0.5 cM), as well as for the haplotypes generated around D5S623 (using two other flanking markers) are 72%. These remain significantly greater than the 50% expected by chance even after correction for multiple testing. Although the IBD approach has advantages for linkage analysis because of its simplicity, all IBD-based methods for quantitative variables may be subject to problems, such as different sibship sizes and missing marker information, and the results require careful interpretation. Most of these problems were overcome by Urbanek *et al.* (1) by using a weighting scheme (20) to account for different sibship sizes and by using several highly polymorphic markers. Although the IBD data indicate strong evidence of linkage to the follistatin gene, the "transmission/disequilibrium test" (TDT) did not show evidence of disease association. What is the explanation for this discrepancy? The power to detect association by using a marker depends on several factors: strength of the linkage disequilibrium between marker and disease, disease predisposing alleles, the recombination fraction between the disease and marker loci, the increase in risk attributable to the particular susceptibility locus under consideration, and the penetrances of the different disease locus genotypes (21). Especially if there are multiple different mutations at the follistatin locus increasing susceptibility to PCOS, failure to find an association is understandable.

How might follistatin be involved in the pathogenesis of PCOS? Follistatin is a high-affinity binding protein that modulates the bioactivity of activin (22, 23). Activin enhances FSH-induced aromatase activity (22, 24), LH binding sites (25), and progesterone production (26, 27) and may play a role in preventing premature luteinization of the ovarian follicle. In the rat model, follistatin modifies FSH action on rat granulosa cells, as evidenced by its inhibition of aromatase activity and inhibin production while enhancing progesterone production (22). Follistatin reverses the enhancing effect of activin on FSH-stimulated steroidogenesis and inhibin production and inhibits activin-induced FSH receptor number (26) and basal inhibin production by granulosa cells (27). Thus, follistatin may modulate granulosa cell function in an autocrine fashion and its mechanism of action is through binding and neutralization of activin action, and it is likely to favor the process of follicular luteinization or atresia. Overexpression of follistatin will therefore be expected to lead to increased ovarian androgen production and reduction in circulating FSH levels, which are features of PCOS.

Previous studies of genetic analysis of PCOS (28–30) have shown inconsistent results, probably because of variations in the definition of the phenotype, study design, and genetic heterogeneity. Waterworth *et al.* (30) evaluated 147 individuals in 14 families. Women were considered affected if they had symptoms of menstrual disturbance and polycystic ovaries on ultrasound. The authors found evidence of linkage with the insulin gene variable number of tandem repeats (VNTR) polymorphism. In addition, there was an association between the insulin VNTR and preferential transmission of the class III allele of the insulin VNTR from heterozygous fathers to PCOS daughters. In the analysis of 14 pedigrees ($n = 142$) by Carey *et al.* (29), the phenotype definition included demonstration of polycystic ovaries by ultrasound and male-pattern baldness. These authors found no evidence of linkage between CYP17 and PCOS, although there was a significant association between the presence of a single base change in the 5′ promoter region (replacement of T by C) at −34 bp from the initiation of translation. The study by Gharani *et al.* (28) included 20 pedigreees (145 members), and evidence of linkage was found between PCOS and the cholesterol side chain cleavage enzyme, CYP11A. Furthermore, they performed an association study of 97 cases and matched controls that demonstrated significant association of a CYP11A 5′-untranslated region pentanucleotide repeat polymorphism with hirsute PCOS subjects. The phenotype definition in the study also included the demonstration of polycystic ovaries by ultrasound.

Overall, the paper by Urbanek *et al.* (1) is an excellent study that uses appropriate family study design, appropriate clinical classification, appropriate selection of loci to test, appropriate statistical analyses, and quite careful and cautious interpretation of statistical results. The study used the radiation hybrid maps (RH) and panels and the detailed STRP maps to identify markers that could be surrogates in a linkage study (affected sibling pairs) for the candidate genes themselves. However, as valuable as those resources have been to human genetics, their utility is rapidly passing. It is currently estimated that within little more than a year from now, there will be a virtually complete reference sequence of the human genome; by 2002, a complete reference sequence should be available. Even now, there is enough sequence in GenBank that the following alternative scenario is possible for many unmapped genes that could be candidates in a study of some complex disorder. With some sequence of the gene, a BLAST search identifies a fully sequenced bacterial artificial chromosome (BAC) with this gene in the middle. A survey of the sequence shows several STR sequences that are long enough they are likely to be polymorphic. Some cosmid and other small clones also overlap with the BAC near the candidate gene, and there are sequence differences that may well be single-nucleotide polymorphisms (SNPs). Primers are designed to amplify these possibly polymorphic sites (STRPs and SNPs) that are molecularly within or immediately adjacent to the candidate gene. Segregation of these loci is examined in the families with the disorder, because even a not-very-polymorphic marker may be segregating in the study families and contribute valuable information on allele sharing. With the proper checks and controls, this approach avoids uncertainties in the RH maps and the uncertain linkage data in the STRP maps; one goes directly to the molecular region of the candidate gene to find markers. Thus, the study by Urbanek *et al.* (1) represents a paradigm for the present, but the use of RH and linkage maps may soon be obsolete.

1. Urbanek, M., Legro, R. S., Driscoll, D. A., Azziz, R., Ehrmann, D. A., Norman, R. J., Strauss, J. F., III, Spielman, R. S. & Dunaif, A. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 8573–8578.
2. Dunaif, A. (1995) *Am. J. Med.* **98,** 33S–39S.
3. Dahlgren, E., Johannson, S. & Lindstedt, G. (1992) *Fertil. Steril.* **57,** 505–513.
4. Conway, G. S., Agrawal, R., Betteridge, D. J. & Jacobs, H. S. (1992) *Clin. Endocrinol.* **37,** 119–125.
5. Sozen, I., Bahtiyar, M. O. & Arici, A. (1996) in *Clinical Practice in Sexuality (Special Issue)*, ed. Walbreck, J. W. (Gordon L. Deal, East Brunswick, NJ), Vol. 11, pp. 49–59.
6. Dunaif, A. A., Segal, K. R., Shelley, D. R., Green, G., Dobrjansky, A. & Licholai, T. (1992) *Diabetes* **41,** 1257–1266.
7. Rosenbaum, D., Haber, R. S. & Dunaif, A. (1993) *Am. J. Physiol.* **264,** E197–E202.
8. Marsden, P. J., Murdoch, A. & Taylor, R. (1994) *Metabolism* **43,** 1536–1542.
9. Williams, G. H. (1994) *Kidney Int.* **46,** 1550–1553.
10. Dummer, M., Greenberg, D. A. & Hodge, S. E. (1992) *Am. J. Hum. Genet.* **51,** 859–870.
11. Goldin, L. R. (1992) *Genet. Epidemiol.* **9,** 61–66.
12. Risch, N. (1990) *Am. J. Hum. Genet.* **46,** 222–228.
13. Risch, N. (1990) *Am. J. Hum. Genet.* **46,** 229–241.
14. Risch, N. (1990) *Am. J. Hum. Genet.* **46,** 242–253.
15. Olson, J. M. (1995) *Am. J. Hum. Genet.* **56,** 788–798.
16. Fulker, D. W. & Cardon, L. R. (1994) *Am. J. Hum. Genet.* **54,** 1092–1103.
17. Kruglyak, L. & Lander, E. S. (1995) *Am. J. Hum. Genet.* **57,** 439–454.
18. Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. & Lathrop, M. (1993) *Nature (London)* **359,** 794–801.
19. Murray, J. C., Buetow , K. H., Weber, J. L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V. C., Sunden, S., Duyk, G. M., *et al.* (1994) *Science* **265,** 2049–2054.
20. Suarez, B. K. & Hodge, S. E. (1979) *Clin. Genet.* **15,** 126–136.
21. Schaid, D. J. & Sommer, S. S. (1994) *Am. J. Hum. Genet.* **55,** 402–409.
22. Xiao, S. & Findlay, J. K. (1991) *Mol. Cell. Endocrinol.* **79,** 99–107.
23. Nakamura, T., Hasegawa, Y., Sugino, K., Kogawa, K., Titani, K. & Sugino, H. (1992) *Biochim. Biophys. Acta* **1135,** 103–109.
24. Xiao, S., Findlay, J. K. & Robertson, D. M. (1990) *Mol. Cell. Endocrinol.* **69,** 1–8.
25. Hutchinson, L. A., Findlay, J. K., de Vos, F. L. & Robertson, D. M. (1987) *Biochem. Biophys. Res. Commun.* **146,** 1405–1412.
26. Xiao, S., Robertson, D. M. & Findlay, J. K. (1992) *Endocrinology* **131,** 1009–1016.
27. Xiao, S., Farnworth, P. G. & Findlay, J. K. (1992) *Endocrinology* **131,** 2365–2370.
28. Gharani, N., Waterworth, D. M., Batty, S., White, D., Gilling-Smith, C., Conway, G. S., McCarthy, M., Franks, S. & Williamson, R. (1997) *Hum. Mol. Genet.* **6,** 397–402.
29. Carey, A. H., Waterworth, D., Patel, K., White, D., Little, J., Novelli, P., Franks, S. & Williamson, R. (1994) *Hum. Mol. Genet.* **3,** 1873–1876.
30. Waterworth, D. M., Bennett, S. T., Gharani, N., McCarthy, M. I., Hague, S., Batty, S., Conway, G. S., White, D., Todd, J. A., Franks, S. & Willamson, R. (1997) *Lancet* **349,** 986–990.
31. Stein, I. F. & Leventhal, M. F. (1935) *Am. J. Obstet. Gynecol.* **29,** 181.