Evolutionary Anthropology 15:160–166 (2006)

Sherlock Holmes and the Empty Cab

It is easy to provide plausible explanations of the main features in evolution. But the little things, which may be quite important, are elusive.

KENNETH M. WEISS

"It has long been an axiom of mine that the little things are infinitely the most important," said Sherlock Holmes to Dr. Watson in *A Case of Identity*.[1] In that story, a Miss Sutherland asks Holmes to find her loving fiancé, Hosmer Angel. Hosmer had disappeared suddenly after making her swear that "Whatever happened, I was to be true . . . pledged to him . . . . strange talk for a wedding morning." Indeed, Angel never arrived at the wedding. He had got into a cab to go there, but disappeared (Fig. 1).

It used to be that science was basically about the big things one could observe with one's own eyes. Aristotle laid down the ground rules in ancient times, and they prevailed for nearly 2,000 years as the dogma about the nature of knowledge of the world. He said that our sensory systems and reasoning powers were provided to enable us to understand the world correctly. But this long era of intuitive truth came crumbling down in the sixteenth and seventeenth centuries. One of the major reasons for this was the discovery that the world could not be seen in its entirety by our senses alone.

ADVANCING THROUGH INSTRUMENTATION

The bombshell dropped in 1609, when Galileo made a serious instrument out of what had originally been a Dutch toy: the telescope. He showed that many things existed that could not be understood by the unaided eye. By undermining the ancient notion that Nature was intuitively discernible, this discovery challenged the Aristotelian dogma and threatened the Church, which had accepted that dogma about the world. The telescope helped usher in the modern era of science based on empiricism rather than intuition.

Galileo showed unsuspected detail in big, distant objects. Then, in 1665, Robert Hooke extended this to the near and small. He opened his *Micrographia,* on use of the microscope and telescope,[2] with a statement of the incremental but empirical progress of science (Fig. 2A). Under a microscope, even the finest human-made objects, such as the point of a fine needle, are "crude-misshapen things" compared to those sculpted by nature itself (Fig. 2B, C).

A fundamental aspect of the new empiricism was the method of induction, the idea that truth emerges from repeated observation. In his experiments on gravity, Galileo had also begun using this criterion.[3] Because clocks were not very accurate, he studied gravity by rolling balls down inclined planes so that their fall took more time and could be measured with a water-clock. He knew there was variation among trials and therefore repeated each experiment 100 times, introducing the idea of replicate experiments. Formal statistical tests were not available, but he knew about measurement error and stressed the convergent results from his multiple trials (of balls rolled down planes and visits to the Inquisition, too!). Repeated observations converged on and revealed the fixed laws of nature.

About two centuries later, science suffered another shock when Darwin and Wallace convinced us that the long-held static view of the world was not true. Instead, life is a dynamic history, driven by inherited causal factors. But these facts could only be inferred in very indirect ways, because neither evolutionary events taking place in the past nor the units of inheritance could be directly observed with the instrumentation available at the time. Evolution and inheritance also forced science into additional new territory. Unlike Newton's laws of physics, which seemed more fixed and deterministic at levels that matter to anyone but a physicist, biological causation seems to be more essentially probabilistic in ways that matter to everyone. Mutation, Mendelian inheritance, natural selection, and genetic drift are inherently probabilistic processes and have major effects on individuals and populations even within a single lifetime. To a great extent, the same is true of the relationships between inherited genotypes and realized phenotypes. Probabilistic causation makes some of the most important questions about life among the most elusive to understand.

FIGURING THE ODDS

For the present, at least, probabilistic causation is approached by a method we can call statistical induction. We state a hypothesis, such as the proposition that a specific genetic cause produces some particular outcome. For example, we may say, "Allele (sequence variant) A has higher selective fitness than does allele B" or "Allele A causes trait D." In broad terms, the way to evaluate assertions such as "A causes D" is to collect a sample of data and determine how often D is found when A is present (Fig. 3A, a and b), then compare that to how often D is present but A, the putative cause, is not (Fig. 3, c and d).

Ken Weiss is Evan Pugh Professor at Penn State University.

DOI 10.1002/evan.20116
Published online in Wiley InterScience (www.interscience.wiley.com).

Figure 1. I saw him get in, but ". . . when the cabman got down from the box and looked, there was no one there." Original illustration by Sidney Paget.[1]

A typical measure of association is the odds ratio (OR = (a/b)/(c/d) = ad/bc). An OR = 1 means that there is no association in the sample between A and D. If OR > 1, then A is positively associated with, or a cause of D; OR < 1 means that A is protective, or reduces the risk of D. Even though causation is not directly observed, an OR > 1 is taken as support for the hypothesis that A causes D. The OR is a measure of the strength of the association. However, the strength of the evidence of association is based on some probabilistic significance test, such as a chi-square *p* value or a confidence interval. That test of the evidence depends on considerations like sample size which are actually unrelated to any causal relation between A and D.

Because of the nature of research funding, most available human examples concern disease. Figure 3B illustrates such results in a whisker-plot form. The top dot is the OR resulting from a study of the association between Alzheimer's disease and a particular allele at the angiotensin-converting enzyme (ACE) gene. The whisker is the 95% confidence limit of the OR estimate. The vertical line is for OR = 1. If the whisker crosses that line, the data are compatible with, or not statistically different from, showing no effect of the putative causal factor, according to our chosen decision criterion (the 95% limits). In this example, the estimated OR is about 1.5 (50% excess risk) and the whisker does not overlap OR = 1.

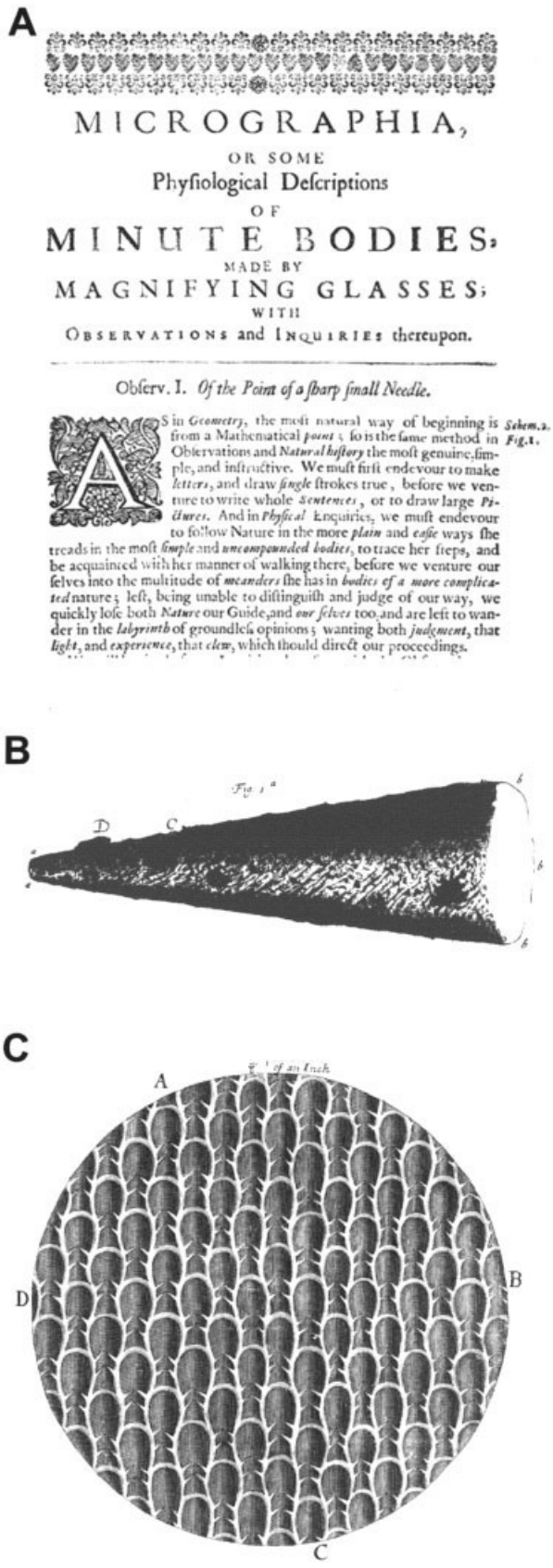

Figure 2. Beginning of *Micrographia:* Robert Hooke gets right to the point. A. First page. B. Magnified man-made needle. C. Natural surface of a kind of seaweed. From Hooke.[2]

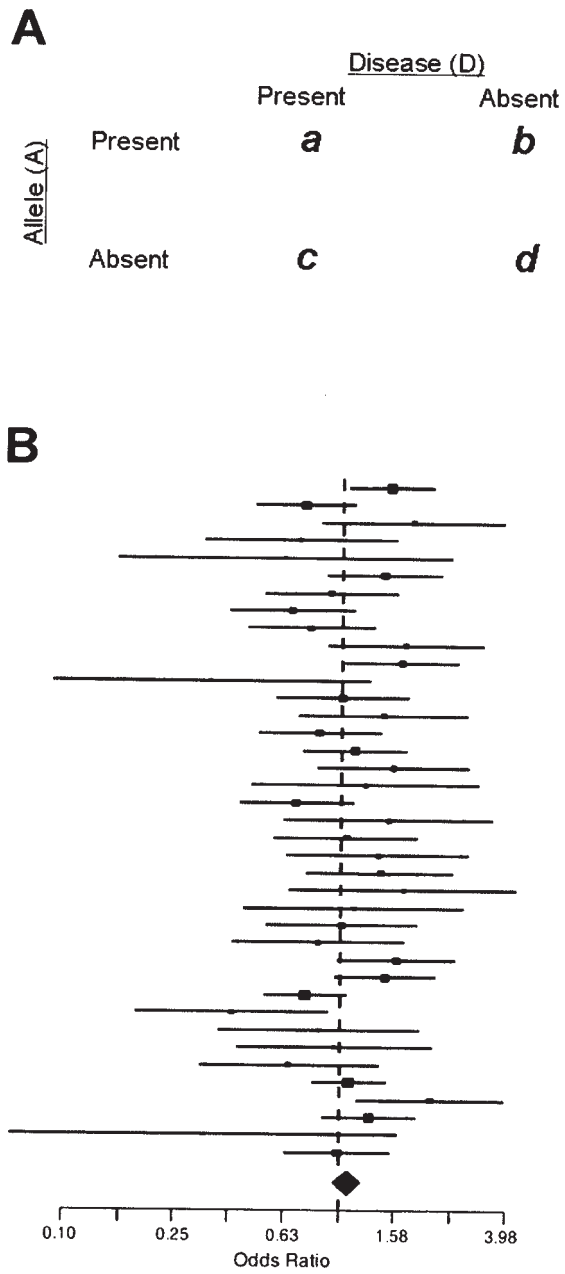

Figure 3. How evidence is evaluated and how it accumulates. A Layout of a genetic case-control study. In this design, the risk is measured by the odds ratio = ad/bc. B. Meta-analysis of the relationship between the presence of a specific nucleotide insertion/deletion allele and Alzheimer's disease. Each row is the result of a different study; the dot gives the OR from that study and the whisker the 95% confidence limits. The vertical dotted line is for OR = 1; that is, there is no association between the allele and the disease. The diamond in the last line is the overall OR for this set of data in a pooled analysis. Data from Lehmann and coworkers.[4]

This is reasonably convincing evidence, and you may be pleased with yourself when you obtain it. But the case is not closed. Because your evidence is only statistical, your friends, unless you've picked a very dull topic, will try to confirm your finding, while your enemies will try to refute it. This is what statistical induction is all about. In Figure 3B, each row but the last gives the results of an additional study of the association between the ACE allele and Alzheimer's disease.

Clearly, each result is different. There is also something remarkable that is typical of replication-association studies. Some studies actually find the reverse effect; in those samples the ACE mutation is found to protect against Alzheimer's disease, at least in the statistical sense of a negative association. In fact, most of the OR estimates are not significantly different from OR = 1. Yet study after study has been done, presumably because of a widespread conviction that there must be some association. Figure 3B is known as a meta-analysis because it not only shows in a single view all the studies that have been done, but also, in the bottom line, gives the result of a pooled analysis of all the data combined. Meta-analysis is becoming quite popular in human genetics and epidemiology because it can achieve much larger samples than can any individual study. In this example, the pooled data show an overall OR of about 1.15, which seems to be barely significant.

So would you say that this ACE allele is associated with Alzheimer's disease? Is it causal or protective? Does the fact that the overall risk is 1.15 mean that every ACE allele carrier is at 15% increased risk? If so, how does a genetic variant cause such a modest increase in the risk of a trait that is not even manifest until after a person has been perfectly healthy for decades? Or does the overall result mean something very different, that 15% of the ACE allele carriers are at 100% risk, while the remainder are at no risk at all? How would you know? Do you conclude that the causal hypothesis is true? Is 1.15 the "true" amount of excess risk and, if so, why did no single study obtain that estimate? Is each separate study result "true?" What do we even mean by "true" in this kind of situation? Should we do more studies? Does it matter how accurately we estimate such values, beyond showing convincingly that they are not equal to 1 and, if not, why?

Despite the apparent simplicity of these questions, abstract Aristotelian reasoning won't answer them. We do not expect each individual observation to be sufficient since, in probabilistic causation, a given person with A might, by chance, not have D. We can't intuit the truth. But probabilistic

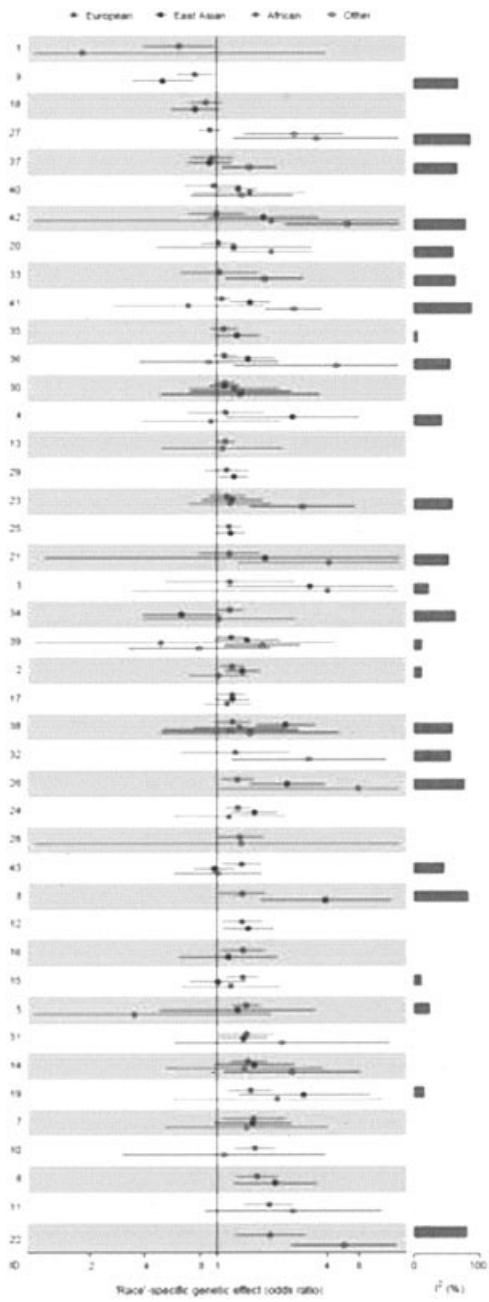

Figure 4. Meta-meta-analysis of 43 disease-gene association results (rows) each studied separately in multiple studies (whisker plots within each row) among Europeans, East Asians, and Africans. (For some traits, there was no data from a given population; others include a pool of "other" populations). Dots represent odds ratio, where OR = 1 means no effect (vertical line). Points to right of the vertical line mean a positive association between genetic variant and disease; points to the left mean a negative (protective) effect. Whiskers are 95% confidence intervals. The length of bars to the far right is proportional to the heterogeneity among regional groups for each gene-trait pair. For example, the top bar represents studies of the effect of the CYP2D6 mutation on risk of lung cancer; the second bar represents the effect of an ACE insertion/deletion polymorphism on diabetic nephropathy. (For data details, see the source[5]). Reprinted with permission from Nature Publishing Group.

causation means that induction does not necessarily provide a clear answer either unless we make additional assumptions that are tantamount to assuming what we want to show in the first place; for example, that there is one true risk that each person experiences. Yet that was the justification for pooling different studies into a single analysis! Significance levels also give a somewhat false impression that we're on solid epistemological ground, because *p*-values are purely subjective, while standard errors are usually "guesstimates" that also are based on ancillary assumptions about the nature of what we want to prove.

These are serious problems in observational science and the issues, with variation in the details, apply to most of quantitative anthropology. (What constitutes good evidence in qualitative anthropology is a very tough question, one I won't dare to answer.) One important question is what our genes do. For example, human populations have occupied different continents for 50,000 or more years. Clearly, regional genetic differences have accumulated across the genome.[6] How important are these differences? Diseases not only constitute traits of practical relevance, but may also be indicators of past natural selection. Studies have been done on the disease-related effects of individual alleles that are found around the world and presumably existed in the ancestral population that expanded out of Africa. If a gene is a gene, and if a gene does something, shouldn't such specific alleles have the same effect wherever they are found?

Figure 4 shows a kind of meta-meta-analysis of 43 different allele-disease associations across diverse genes and diseases. Each row shows the results of three separate meta-analyses, each based on multiple studies of the same gene-disease association within Africans, Europeans, and East Asians, respectively. For most diseases, the three studies' overall OR values (dots in the figure) are fairly similar for the three continental groups. This is reassuring from a genetic-mechanism point of view. But there is also substantial variation and uncertainty, especially as shown by the error bars, which, remember, are for meta-analyses, not individual studies. The individual studies will also have large standard errors, as in Figure 3. Moreover, we can't rule out heterogeneity among individuals within each study. This variation is not trivial; it is, in fact, at least a bit

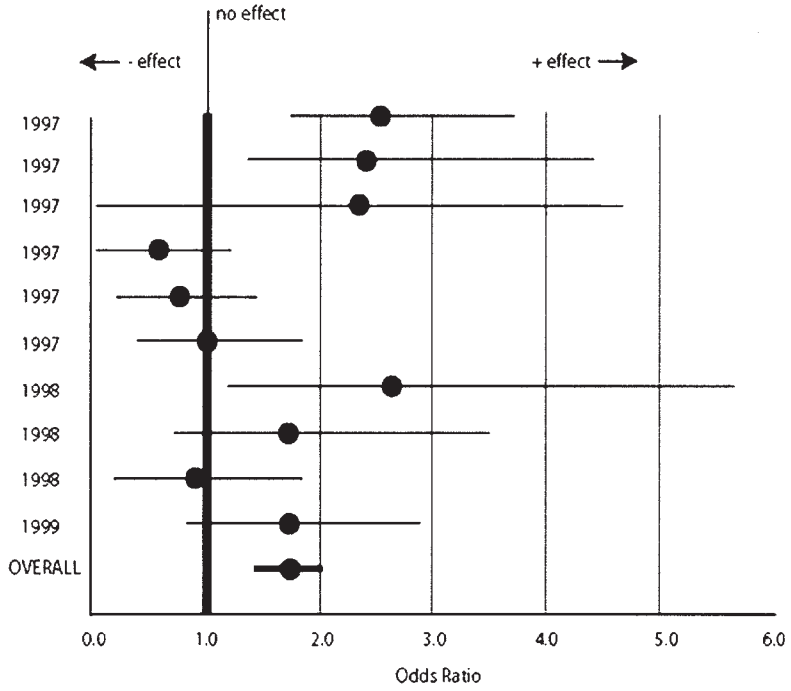

Figure 5. Meta-analysis of studies of the efficacy of Mystery Factor X (answer given in NOTES). Data from Bem, Palmer, and Broughton;[7] see Matthews.[8]

unsettling. Indeed, in some cases the meta-result for a population on one continent shows the opposite effect of the same mutation for the same trait in other continental populations. What makes a difference large or small in this kind of analysis is a matter of judgment. Also, because data are limited, the East Asians include Indochinese and Philippine samples, while Africans include African Americans; the latter and possibly the former have substantial European admixture, and thus water down the comparison the authors sought to make.

But does this direct genetic evidence indicate that a gene is a gene no matter where it is found? Actually, we know that the same allele need not always have the same effect. Engineering the same human mutation in the homologous gene of a mouse routinely has different effects from those seen in humans. Further, those effects usually vary among mouse strains. Such differences may sometimes be due to environment; after all, we don't subsist on Purina mouse chow. But the main reason for the differences is that while primates and mice have basically the same genes, these differ in detail between the two species. Thus, a given mutation finds itself in a substantially different genomic background in mice as compared to that in humans. Indeed, even if it is to a lesser extent, human genomes vary enough, quite aside from environmental differences, that, except in twins, the same allele is never found in exactly the same genomic background. So are the variable results due to this variation, sample variation, environmental variation, or an inherent probabilistic nature of what the gene is doing?

The problem is actually more subtle even than this. Accumulating evidence by induction gradually to refine confidence in a Truth belies unstated biases that have little to do with the data themselves. There are no formal criteria for reaching consensus, as there are in mathematical proofs, and the believability of "A causes D" involves many unstated factors. These include the authors' commitments to their hypotheses, preconceptions that affect how data are collected, the need to obtain grant support, flattery by the media for simplified "findings," publishing to earn tenure, the difficulty of publishing negative results, and so on. Investigators of human "racial" variation almost always work with an agenda based on their sociopolitical view of that touchy subject.

How many negative studies would it take to cause even an objective person to abandon belief in a hypothesis he or she advocated? By the time half the studies in Figure 3 were done, why would anyone want to look at the ACE situation again? Does it imply a commitment to the importance of this gene? Or does it imply that a positive finding will lead to future grant support? How heavily are you likely to criticize, or how easily believe, a study result that does not fit your preconceptions?

Depending on your world view, questions like these may or may not sound too postmodern for your taste. But look at the meta-analysis in Figure 5. This shows an overall OR of about 1.7, a result much stronger than those of most human-disease genetic meta-analyses such as the ones presented in Figures 3 and 4. So we should readily accept that strong result. Any takers?

"OH, IT DRIVES ME HALF MAD TO THINK OF!"[1]

For many things we'd like to understand, even erroneous speculation does no harm, but if false facts are widely accepted, understanding can be set back for a long time as, for example, was the case with Piltdown in anthropology. The reason we evolved opposable thumbs probably does have to do with tool use rather than, say, hitchhiking. Somewhat more speculative, though still harmless, are scenarios explaining our small canines. Do they represent a shift of defense from teeth to weapons? A display threat shifted from fang-bearing to face-paint, feather head-dresses, and vocalization? A hominid change to an omnivorous diet? But when we attempt to explain things currently having potential health or social consequences, little differences in scientific interpretation can make all the difference to people's lives, because if they are assumed to be important then medical research investment, therapy, and behavior can follow, sometimes falling victim to the "law" of unintended consequences.

Well-intended therapies can turn out to have net harmful effects, as may be the case when cardiovascular disease increases under hormone replacement for menopausal symptoms, kidney stones ensue from dietary calcium supplements designed to prevent osteoporosis, or birth defects arise from the use of thalidomide for morning sickness.

Relatively simple single-gene causation, often based on evolutionary scenarios, has recently been attributed to several potentially important aspects of human behavior. A few examples are alleles that, based on early reports, have been said to be responsible for human brain size or intelligence, aggressiveness, response to stress, and novelty-seeking. Less well known than the highly publicized initial findings are meta-analyses that fail to replicate or seriously mute the original finding.[9,10–13,14] There is no fault involved except, perhaps, unrealistic claims or expectations, which can detract attention from understanding the subtle way evolution usually works. But cases where multiple samples can be studied are the easy ones. In most areas of anthropology, we get only one direct look at a question. Each archeological site is unique, and our biological evolution happened only once. Having to deal with nonreplicable evidence is like throwing a dart at Figure 3 and having the study the dart hits be your entire body of evidence. But biomedical genetics can attempt to replicate studies; it serves as a kind of instrumentation that shows some of the problems we face in dealing with probabilistic causation. Thus, if it's easy to know the main features of what a gene does, the devil, along with important consequences, is in the proverbial details. The news isn't all bad, by any means, because most confidence intervals in Figure 3 overlap. Most studies are, in that sense, compatible with each other—if you accept the assumptions underlying the confidence intervals. But you might come to very different conclusions if you had only one study to work with, since we tend to interpret things in terms of the OR itself, not its confidence limits.

Greatly improved DNA-sequencing instrumentation has brought us ever closer to what seem like the primary causal elements of life. But the little things this reveals show uncertainties that can drive us half mad, as Mr. Angel's disappearance did Miss Sutherland. Perhaps evolutionary biology and genetics still haven't got close enough to all of the relevant variables. But when they do, we may see that the point of interest, rather than converging on the smooth truth we expected, like Hooke's needle becomes rougher and shows things not otherwise suspected. Hooke observed that human artifacts, "even in those most neat," would appear less beautiful if seen close up, but he thought "the works of Nature, the deepest Discoveries show us the greatest Excellencies." His examples were many, such as the perfect beauty of crystals like snowflakes and a fly's eye. But while evolution is also part of nature it is such a "crude-misshapen" process that it blurs when looked at too closely.

These are serious and widespread problems, but they are controversial because they challenge many aspects of current practice.[15,16] If the replication of studies does not lead us asymptotically to a single truth, the inductive core of empiricism is undermined. Science is not designed for the study of uniqueness. What a science of uniqueness might look like, if it is even possible to develop one, perhaps can be discovered only by a young person whose brain cells are unpolluted by our current methodological preconceptions.

Like Miss Sutherland's fiancé, meta-analysis shows how something that seems to be there on one look can disappear on a second look. Sherlock Holmes solved his case by old-fashioned intuition rather than instrumentation: Mr. Angel had stepped into the cab, shut the door, and slipped unseen out the other side. His disappearing act showed that he was no angel. Indeed, he wasn't even a fiancé. He was Miss Sutherland's stepfather in disguise, playing the disappearing fiancé so she would not really marry someone and remove her annuity from his control. His dual identity was discovered when Holmes used his own sharp eyes to see that two documents had been printed on the same typewriter. Poor Miss Sutherland had been destined for spinsterhood.

In studying evolution, perhaps improved instrumentation will eventually lead to a more satisfying kind of causation than we have now but, as Hooke might suggest, we do not yet have the right lens before our eyes. On the other hand, when the problem we are struggling with is conceptual rather than physical, the instrument we are not applying intensely enough is the one behind our eyes.

NOTES

The studies in Figure 5 are on the efficacy of extrasensory perception.[7,8] I welcome comments on this column. I have a feedback and supplemental material page at http://www.anthro.psu.edu/weiss_lab/index.html. I thank Anne Buchanan, Sam Sholtis, and John Fleagle for critically reading this manuscript.

REFERENCES

Quotes from online material can be found by using Find in a browser.

1 Doyle AC. 1891. A case of identity. Strand Magazine.

2 Hooke R. 1665. Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses with observations and inquiries thereupon. London: Martyn and Allestry. Available online at http://sherlock-holmes.classic-literature.co.uk/a-case-of-identity/

3 Galilei G. 1638. Dialogues concerning two new sciences. Leiden: Elzevir (English Dover edition, 1954). Available online from Project Gutenberg at http://www.gutenberg.org/etext/15491

4 Lehmann DJ, Cortina-Borja M, Warden DR, Smith AD, Sleegers K, Prince JA, van Duijn CM, Kehoe PG. 2005. Large meta-analysis establishes the ACE insertion-deletion polymorphism as a marker of Alzheimer's disease. Am J Epidemiol 162:305–317.

5 Ioannidis JP, Ntzani EE, Trikalinos TA. 2004. "Racial" differences in genetic effects for complex diseases. Nat Genet 36:1312–1318.

6 Weiss KM, Fullerton SM. 2006. Racing around, getting nowhere. Evol Anthropol 14:165–169.

7 Bem D, Palmer J, Broughton R. 2001. Updating the Ganzfeld database: a victim of its own success? J Parapsychol 65:207–218.

8 Matthews R. 2004. Opposites detract. New Sci 181:39–41.

9 Kluger AN, Siegfried Z, Ebstein RP. 2002. A meta-analysis of the association between DRD4 polymorphism and novelty seeking. Mol Psychiatry 7:712–717.

10 Preisig M, Ferrero F, Malafosse A. 2005. Monoamine oxidase a and tryptophan hydroxylase gene polymorphisms: are they associated with bipolar disorder? Am J Pharmacogenomics 5:45–52.

11 Schinka JA, Letsch EA, Crawford FC. 2002. DRD4 and novelty seeking: results of meta-analyses. Am J Med Genet 114:643–648.

12 Tochigi M, Hibino H, Otowa T, Kato C, Marui T, Ohtani T, Umekage T, Kato N, Sasaki T. 2006. Association between dopamine D4 receptor (DRD4) exon III polymorphism and neuroticism in the Japanese population. Neurosci Lett 398:333–336.

13 Tochigi M, Hibino H, Otowa T, Ohtani T, Ebisawa T, Kato N, Sasaki T. 2006. No association of 5-HT(2C), 5-HT(6), and tryptophan hydroxylase-1 gene polymorphisms with personality traits in the Japanese population. Neurosci Lett.

14 Zammit S, Owen MJ. 2006. Stressful life events, 5-HTT genotype and risk of depression. Br J Psychiatry 188:199–201.

15 Buchanan AV, Weiss KM, Fullerton SM. 2006. Dissecting complex disease: the quest for the Philosopher's Stone? Int J Epidemiol 35: 562–571, 593–596.

16 Buchanan AV, Weiss KM, Fullerton SM. 2006. On stones, wands, and promises. Int J Epidemiol. In press.

Books Received

- Cachel, S. (2006) Primate and Human Evolution. xvii+469 pp. New York: Cambridge University Press. ISBN 0-52182-942-9 $120.00 (cloth).
- Lieberman, P. (2006) Toward an Evolutionary Biology of Language. ix+427 pp. Cambridge: Harvard University Press. ISBN 0-67402-184-3 $49.95 (cloth)
- Valentine, J.W. (2006) On the Origin of Phyla. xxiv+614 pp. Chicago: University of Chicago Press. ISBN 0-22684-548-6 $35.00 (paper).
- Rampton, B. (2006) Language in Late Modernity: Interaction in an Urban School. xviii+443 pp. Cambridge: Cambridge University Press. ISBN 0-52181-263-1 $95.00 (cloth).
- Madrigal, L. (2006) Human Biology of Afro-Carribean Populations. xv + 251 pp. Cambridge: Cambridge University Press. ISBN 0-521-81931-8 $125.00 (cloth).
- Webb, S. (2006) The First Boat People. xvii + 318 pp. Cambridge: Cambridge University Press. ISBN 0-521-85656-6 $120.00 (cloth).
- Carrano, M. T., Gaudin, T. J., Blob, R. W., and Wible, J. R. Amniote Paleobiology. vi + 547 pp. Chicago: University of Chicago Press. ISBN 0-226-09477-4 $95.00 (cloth). ISBN 0-226-09478-2 $40.00 (paper).
- Nazarea, V. D. (1998) Cultural Memory and Biodiversity. xiv + 189 pp. Tucson: University of Arizona Press. ISBN 0-81652547-1 (paper) $24.95.
- Trigger, B. W. (2006) A History of Archaeological Thought, 2nd Edition. xx + 710 pp. Cambridge: Cambridge University Press. ISBN 0-521-60049-9 $31.99 (paper).
- Hawkes, K. and Paine, R. R., eds. (2006) The Evolution of Human Life History. xiii + 505 pp. Sante Fe: School of American Research Press. ISBN 1-930618-72-7 $34.95 (paper).
- Meldrum, J. (2006) Sasquatch: Legend Meets Science. 297 pp. New York: TOR/Forge Press. ISBN 0-765-31216-6 $27.95 (cloth).
- Chamberlain, A. T. (2006) Demography in Archaeology. xix + 235 pp. Cambridge: Cambridge University Press. ISBN 0-521-59651-3 $29.99 (paper).
- Gomez, J. C. (2004) Apes, Monkeys, Children, and the Growth of Mind. xi + 340 pp. Cambridge: Harvard University Press. ISBN 0-674-02239-4 $19.95 (paper).
- Hall, B. K. and Olson, W. M. eds. (2003) Keywords and Concepts in Evolutionary Developmental Biology. xvi + 476 pp. Cambridge: Harvard University Press. ISBN 0-674-02240-8 $29.95 (paper).
- Platek, S. M. and Shackelford, T. K., eds. (2006) Female Infidelity and Paternal Uncertainty: Evolutionary Perspectives on Male Anti-Cuckoldry Tactics. viii + 248 pp. Cambridge: Cambridge University Press. ISBN 0-521-60734-5 $55.00 (paper).
- Dugatkin, L. A. (2006) The Altruism Equation: Seven Scientists Search for the Origins of Goodness. xi + 188 pp. Princeton: Princeton University Press. ISBN 0-691-12590-2 $24.95 (cloth).
- Scientific American, ed. (2006) Evolution: A Scientific American Reader. vii + 355 pp. Chicago: University of Chicago Press. ISBN 0-226-74269-5 $22.00 (paper).
- Lloyd, E. A. (2006) The Case of the Female Orgasm. 311 pp. Cambridge: Harvard University Press. ISBN 0-674-02246-7 $17.95 (paper).