Human genome editing as a tool to establish causality

Fyodor D. Urnov[1]
Sangamo BioSciences, Richmond, CA 94804

Gentle reader, are you able to drink whole milk? If yes, you have a single base pair in your genome to thank. Certain people can consume milk as adults because of agricultural practice–imposed selection for that ability, which has led to an increase in the frequency of a fascinating allele of the lactase gene. The lactase coding region has remained unchanged, and instead, the enhancer of the gene, which regulates the timing of its expression and lies 14,000 bp upstream of the transcription start site, has acquired a SNP that allows it to be expressed in adults (1). This theme—a variant in a noncoding *cis*-regulatory region that exerts a major effect on a human—is the focus of an impressive study by Ochiai et al. in PNAS (2). Particularly noteworthy is the use of genome editing with engineered nucleases to establish causality between a noncoding genomic variant and a specific condition.

The facts of the patient case history described by Ochiai and colleagues are deeply affecting. Two healthy parents have a healthy daughter and then have a son struck with a severe disease: premature chromatid separation (PCS) syndrome, which is caused by loss-of-function mutations in a gene encoding Bubr1, a protein component of the spindle assembly checkpoint. The boy was treated for cancer, which is in remission; he is nonverbal and suffers from seizures. Both parents appeared to be carriers for the disease-causing allele and were planning to have a third child. Given the impact made by prenatal diagnostics on the incidence of severe monogenic disease in affected populations (3, 4), it would seem that the "standard of care" path forward was clear. The challenge, however, was that no coding mutation in Bubr1 was found in either of the parents, suggesting that either the boy's symptoms were not caused by a lesion in Bubr1 or that the causative mutation is a noncoding one. Ochiai et al. resolve this dilemma, and this not only allowed this family to have a third child, a healthy boy, but provided a valuable lesson for genomicists.

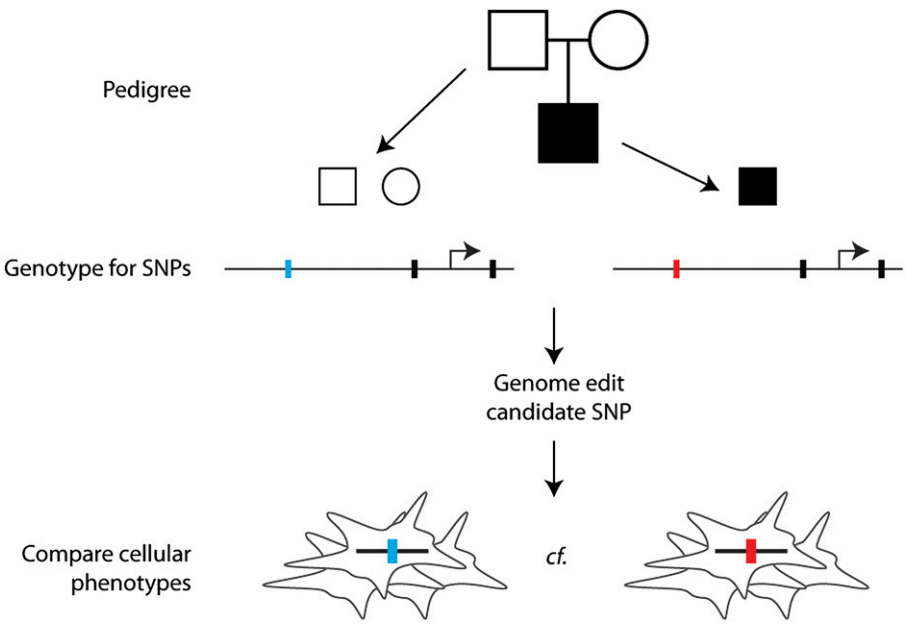

Fig. 1. An outline of the experimental scheme used by Ochiai et al.

"Post hoc, ergo propter hoc": If there is such a thing as hell for people who study the effects of human genetic variation on phenotype, then these words would be written over the gates to that place (in lieu of Dante's "lasciate ogni speranza"). Translated from the Latin as "after something, therefore because of it," this well-known "coincidental correlation" logical fallacy points out that determining whether something (e.g., a SNP) causes a process of interest (e.g., a disease) or is merely correlated with it is a major unmet challenge. The words malaria and influenza remind us all of this with respect to human health. When the diseases were given their names, it was believed the former was caused by bad air (mal'aria in Italian) emitted by swamps, and the latter was due to the influence (influenza, also in Italian) of the planets on the health of people.

The study of the genetic basis of monogenic disease faces this correlation vs. causality problem daily. Following pioneering work from the Shendure laboratory (5), the impact of next-generation sequencing in this setting has been striking and is exemplified by work from Worthey et al., who used it to identify a coding variant in XIAP as causative of severe gastrointestinal symptoms in a pediatric patient and then relied on this genetic diagnosis to guide the care of this patient to success (6). Genomics, however, seldom yields a single coding variant as causative: more often than not, even after application of the most sophisticated set of filters, there remains more than one variant left standing. The predicament with respect to noncoding variation, where one cannot ask whether a given variant causes a major or minor effect on protein function, is even more dire. More than 90% of all of the variants identified in genome-wide association studies are noncoding. What do they do? In addition to the aforementioned noncoding SNP affecting the lactase enhancer, a recent example comes from the laboratory of Orkin, which traced the genetic signature of fetal hemoglobin levels

Author contributions: F.D.U. wrote the paper.

Conflict of interest statement: F.D.U. is an employee of Sangamo Biosciences, Inc.

See companion article on page 1461.

[1]E-mail: furnov@sangamo.com.

in adults to a SNP in a erythroid-specific enhancer of the Bcl11a gene (7). These two examples of a clear "from SNP to mechanism" understanding are the exception, however, and not the rule, in human genomics.

Ochiai and colleagues decisively cut through the Gordian knot of noncoding variation–imposed confusion. The parents of the affected boy carry, and the child himself is homozygous for, an extended haplotype of the Bubr1 gene. This genomic stretch was sequenced, and after common variants were excluded, a single SNP, residing 44 kb upstream of the gene's start site, was identified as potentially causative. Or did it merely correlate with the symptoms? The authors turned to genome editing to answer this question.

Just as developments in next-generation sequencing have changed the practice of medical genomics, the craft of human somatic cell genetics has dramatically changed in the last 8 y. This is due to the development of genome editing with engineered nucleases; partially as a result of early work by the Jasin (8), Klug (9), Pabo (10), and Carroll (11) laboratories, researchers in laboratories all over the world can now alter the sequence of endogenous loci in human cells. All of the principles and approaches of editing were initially developed, elaborated, and validated using zinc finger nucleases (ZFNs) (12); two more recent platforms for genome editing, transcription activator-like (TAL) effector nucleases (13–15), and the RNA-guided endonuclease Cas9 (16, 17), have adopted these principles and approaches wholesale in pursuit of the cause of rational genomics.

In genome editing, one uses an engineered nuclease to induce a double-strand break in the locus of interest. The break can be repaired using an exogenous template to transfer novel information into the genome. In its first application to an endogenous locus in human cells (18), 20% of all copies of the IL2Rγ gene acquired a novel, investigator-defined SNP, in a single step and without selection. A recent example of the use of such gene correction to understand how known disease-causing variants exert their effect is a study that used ZFNs to generate a panel of isogenic induced pluripotent stem cell lines bearing a Parkinson's disease–related mutation in the α-synuclein gene (19) and then used those lines to reveal the mechanism through which these mutations contribute to disease pathogenesis (20).

Similarly, to determine whether the noncoding SNP somehow affects Bubr1 expression from its location 44 kb upstream, Ochiai and colleagues use genome editing with TAL effector nucleases (14) to generate a panel of cells bearing the WT allele of Bubr1, homozygous for the candidate mutant allele, and, as an important control, cells in which editing was used to regenerate the WT allele. These cells were then compared with each other (Fig. 1). De novo introduction of the noncoding variant lowered levels of Bubr1 mRNA and protein. In agreement with the role of Bubr1 in the mitotic spindle function, genome-edited cells bearing this "transcriptional hypomorphic" allele of Bubr1 exhibited a mild but clear impairment in their mitotic function as gauged by karyotyping, direct measurement of premature chromatid separation, and sensitivity to an inhibitor of the mitotic spindle assembly checkpoint. Are these unique phenotypes an artifact of the genome editing procedure itself? The evidence clearly argues against this: single cell–derived clones that have gone through genome editing to reengineer the WT allele were shown to be identical to parental cells. Thus, the noncoding variant is not merely correlated with the disease-causing haplotype but creates a hypomorphic variant of Bubr1, likely by interference with function of a transcriptional enhancer. In support of that line of reasoning, chromosome conformation capture analysis showed an interaction between the SNP-bearing region and the promoter of Bubr1; furthermore, a signature of active chromatin was seen in the vicinity of the causative SNP as gauged by comparison of SNP location to data from the Encyclopedia of DNA Elements (ENCODE) consortium.

This result had an immediate impact on the lives of the affected family: the parents chose to perform amniocentesis during the course of the third pregnancy, and the newly established disease-causing locus was genotyped in the fetus. It was found to be heterozygous, and because the allele is fully recessive to WT, a healthy baby boy was born after an otherwise uneventful pregnancy. Thus, in its use of every genomic tool available to guide patient care and family planning, the paper by Ochiai and colleagues is a paragon of how medical genomics can affect public health.

The broader implications of this work come from the juxtaposition of two recent developments. First, the potential impact of noncoding variation on regulatory DNA can be now rapidly assessed in the context of extensive, high-quality epigenomic maps provided by the ENCODE consortium (www.encode-roadmap.org): for instance, work from the Stamatoyannopoulos laboratory has shown that a remarkable 57% of GWAS-identified SNPs fall into active regulatory elements (21). Second, genome editing is applicable to any locus in the genome using no fewer than three nuclease platforms (22). Of specific relevance to correction of SNPs, Davis and colleagues demonstrated the utility of oligonucleotides as donor templates to generate novel allelic variants of endogenous loci in a single step and without selection (23).

David Hilbert, father of the enormously influential list of problems for mathematicians to solve, once said "Wir müssen wissen. Wir werden wissen" (we must know; we will know). The paper from Ochiai et al. and the advances in technology that it relies on give one that level of optimism for the future of medical genomics.

1 Enattah NS, et al. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30(2):233–237.

2 Ochiai H, et al. (2014) TALEN-mediated single-base-pair editing identification of an intergenic mutation upstream of *BUB1B* as causative of PCS (MVA) syndrome. *Proc Natl Acad Sci USA* 111:1461–1466.

3 Bach G, Tomczak J, Risch N, Ekstein J (2001) Tay-Sachs screening in the Jewish Ashkenazi population: DNA testing is the preferred procedure. *Am J Med Genet* 99(1):70–75.

4 Cao A, Kan YW (2013) The prevention of thalassemia. *Cold Spring Harb Perspect Med* 3(2):a011775.

5 Ng SB, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 42(1):30–35.

6 Worthey EA, et al. (2011) Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 13(3):255–262.

7 Bauer DE, et al. (2013) An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* 342(6155):253–257.

8 Rouet P, Smih F, Jasin M (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* 14(12):8096–8106.

9 Miller J, McLachlan AD, Klug A (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J* 4(6):1609–1614.

10 Pavletich NP, Pabo CO (1991) Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 A. *Science* 252(5007):809–817.

11 Bibikova M, et al. (2001) Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol Cell Biol* 21(1):289–297.

12 Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* 11(9):636–646.

13 Boch J, et al. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* 326(5959):1509–1512.

14 Miller JC, et al. (2011) A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* 29(2):143–148.

15 Joung JK, Sander JD (2013) TALENs: A widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol* 14(1):49–55.

16 Wiedenheft B, Sternberg SH, Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482(7385):331–338.

17 Ran FA, et al. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 8(11):2281–2308.

18 Urnov FD, et al. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* 435(7042):646–651.

19 Soldner F, et al. (2011) Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* 146(2):318–331.

20 Ryan SD, et al. (2013) Isogenic human iPSC Parkinson's model shows nitrosative stress-induced dysfunction in MEF2-PGC1α transcription. *Cell* 155(6):1351–1364.

21 Maurano MT, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190–1195.

22 Gaj T, Gersbach CA, Barbas CF, 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol* 31(7):397–405.

23 Chen F, et al. (2011) High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat Methods* 8(9):753–755.