Identifying Chronic Obstructive Pulmonary Disease Genes: Shining the Light on Dark DNA

The discovery that mutations in the alpha-1 antitrypsin gene predisposed subjects to emphysema, combined with experimental models in the 1960s showing that instillation of elastases resulted in emphysema, led to the elastase-antielastase hypothesis and ushered in the modern age of emphysema pathogenesis. With subtle but important variations on the theme, it remains central to our understanding of emphysema, an important component of chronic obstructive pulmonary disease (COPD). As genetic techniques and computational sciences advanced our ability to perform genome-wide association studies (GWASs) (1, 2) and more recently whole-genome sequencing (3, 4), our expectation was that we would find the critical genes responsible for major diseases, define disease-causing molecular pathways, and develop interventions to improve care, if not cure the diseases. So, where are the cures?

Although GWASs have successfully identified numerous genetic loci associated with COPD and its associated phenotypes, determining their biological relevance has remained a challenge (1, 2). A few genes that have been implicated in GWASs, such as *HHIP*, are supported by additional *in vitro* and *in vivo* evidence, although they still lack a clear clinical utility, and the biological link for many loci remains unclear (5). This is explained by the complex genetic architecture that underlies chronic diseases like COPD, with contributions from numerous single-nucleotide variants (SNVs) and structural variants that occur at both high and low frequency in the population. This architecture can confound the interpretation of results of classic GWASs that have primarily focused on SNVs that occur with relatively high frequency in the population, for several reasons. First, variants identified in these studies often "tag" other variants due to complex patterns of linkage disequilibrium, which means that the SNV with the lowest P value does not necessarily represent the causal variant (6). Furthermore, many of the variants that have been identified lie in noncoding "dark DNA," and although we have begun to recognize that these variants can affect the function of genes both near and far, we are limited in our ability to predict these effects.

One approach to overcome this problem is to identify loci in GWASs that fall below strict multiple testing correction thresholds but have known functional effects. This approach is supported by our knowledge that many traits with a complex pattern of inheritance have a large number of contributing variants. As an extreme example, height has been associated with over 700 unique loci, but these variants only explain approximately one-quarter of the heritability of that trait (7). However, by progressively including variants at a lower threshold, we can explain a greater portion of that heritability (8). In COPD, where it has been shown that the most frequently replicated loci that reach genome-wide significance in GWASs only account for a small portion of the heritability of the disease, this same effect is likely to play some role (9). Thus, using evidence of functional relevance from other datasets to guide our search to determine which of the many GWAS results that do not reach genome-wide significance may be biologically relevant can be a powerful tool. Some recent examples of this approach in COPD include the integration of gene expression and GWAS results (10), and the combination of mouse and human GWASs (11). These types of approaches can be further expanded by including additional "big" datasets from other omics studies and asking well-directed questions of them (12).

In this issue of the *Journal*, Boueiz and colleagues (pp. 388–398) present a novel integrative method to identify variants that fall below multiple testing thresholds (in this case, $P < 5 \times 10^{-5}$) but may still be biologically relevant to the COPD-related phenotype emphysema apico–basal distribution (13). They combine GWAS data with expression quantitative trait loci (eQTL) from the publicly available Genotype Tissue Expression database, as well as newly collected eQTL from individuals included in the COPDGene (Genetic Epidemiology of COPD) cohort, which formed a portion of their GWAS study population. They aimed to identify functional loci that affect gene expression. They also looked at overlap between their GWAS results and epigenomic markers using data available from the Roadmap Epigenomics project, which gave them loci that, based on their location in regulatory elements, are also likely to affect gene expression. Because both eQTL and epigenomic data were collected from multiple cell types, they are also able to offer evidence that most of these loci exert their effects in multiple cell types. They then go on to identify four loci that are supported by both regulatory and eQTL data, finally supporting the functional role of one of these variants with additional studies *in vitro* and a validating gene expression study to support the role of the associated candidate gene *ACVR1B*. This activin receptor, in the transforming growth factor-β signaling pathway, raises interesting hypotheses as to how it might relate to emphysema through either matrix production or inflammation. The fact that expression of *ACVR1B* in T cells was implicated further suggests a role in immunity and inflammation.

Importantly, throughout their work, Boueiz and colleagues thoroughly and thoughtfully used bioinformatics tools and mathematical methods to support their questions. Although integrative approaches such as those presented by Boueiz and colleagues offer a method of identifying genetic loci that may be biologically relevant to disease processes, they also rely on inherent assumptions. For example, in this study, overlap was determined with an arbitrary window size and all regulatory regions identified in the Roadmap Epigenomics Project were treated with equal weight. Furthermore, just over half of the emphysema apico–basal distribution loci they identified occurred in a regulatory element, meaning that at least another half of these loci will require exploration with other methods. These assumptions, then, rather than emphasizing major problems with the approach, demonstrate the importance of additional carefully designed bioinformatics studies that ask biologically or clinically supported questions.

In an era in which large omics datasets are being collected at an increasingly rapid pace, it is more important than ever to think about how these data can be integrated and interrogated in thoughtful ways to identify biologically or clinically relevant findings. Achieving this goal will require continued collection and integration of large datasets from multiple sources, improved bioinformatics techniques to analyze the data, and functional assessments of discoveries at the bench. Ultimately, leveraging all of these "big data" to shed light on clinically actionable discoveries currently hidden in the "dark DNA" will significantly contribute to future medical advances, allowing us to implement precision medicine, and bring us closer to the original promise of novel interventions for genetically complex diseases like COPD and perhaps even one day to a cure. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Josiah E. Radder, M.D., Ph.D.
Steven D. Shapiro, M.D.
Department of Medicine
University of Pittsburgh Medical Center and University of Pittsburgh
Pittsburgh, Pennsylvania

References

1. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, *et al.*; Understanding Society Scientific Group; Geisinger-Regeneron DiscovEHR Collaboration. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* 2017;49:416–425.
2. Cho MH, Castaldi PJ, Hersch CP, Hobbs BD, Barr RG, Tal-Singer R, *et al.*; NETT Genetics, ECLIPSE, and COPDGene Investigators. A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *Am J Respir Crit Care Med* 2015;192:559–569.
3. Radder JE, Zhang Y, Gregory AD, Yu S, Kelly NJ, Leader JK, *et al.* Extreme trait whole-genome sequencing identifies PTPRO as a novel candidate gene in emphysema with severe airflow obstruction. *Am J Respir Crit Care Med* 2017;196:159–171.
4. Prokopenko D, Sakornsakolpat P, Loehlein Fier H, Qiao D, Parker MM, McDonald MN, *et al.*; COPDGene Investigators, NHLBI TOPMed Investigators. Whole genome sequencing in severe chronic obstructive pulmonary disease. *Am J Respir Cell Mol Biol* 2018;59:614–622.
5. Lao T, Glass K, Qiu W, Polverino F, Gupta K, Morrow J, *et al.* Haploinsufficiency of Hedgehog interacting protein causes increased emphysema induced by cigarette smoke through network rewiring. *Genome Med* 2015;7:12.
6. Orozco G, Barrett JC, Zeggini E. Synthetic associations in the context of genome-wide association scan signals. *Hum Mol Genet* 2010;19: R137–R144.
7. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, *et al.*; GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum Mol Genet* 2018;27:3641–3649.
8. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, *et al.*; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014;46: 1173–1186.
9. Zhou JJ, Cho MH, Castaldi PJ, Hersh CP, Silverman EK, Laird NM. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am J Respir Crit Care Med* 2013;188: 941–947.
10. Sakornsakolpat P, Morrow JD, Castaldi PJ, Hersh CP, Bossé Y, Silverman EK, *et al.* Integrative genomics identifies new genes associated with severe COPD and emphysema. *Respir Res* 2018; 19:46.
11. Radder JE, Gregory AD, Leme AS, Cho MH, Chu Y, Kelly NJ, *et al.* Variable susceptibility to cigarette smoke-induced emphysema in 34 inbred strains of mice implicates Abi3bp in emphysema susceptibility. *Am J Respir Cell Mol Biol* 2017;57:367–375.
12. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014; 15:162.
13. Boueiz A, Pham B, Chase R, Lamb A, Lee S, Naing ZZC, *et al.*; Genetic Epidemiology of COPD (COPDGene) Investigators. Integrative genomics analysis identifies *ACVR1B* as a candidate causal gene of emphysema distribution. *Am J Respir Cell Mol Biol* 2019;60:388–398.