Exome sequencing makes medical genomics a reality

Leslie G Biesecker

Massively parallel sequencing of the exomes of four individuals with Miller syndrome, combined with filtering to exclude benign and unrelated variants, has identified causative mutations in *DHODH*. This approach will accelerate discovery of the genetic bases of hundreds of other rare mendelian disorders.

The genes underlying mendelian disorders have for the past several decades been identified through positional cloning, a process of meiotic mapping, physical mapping and candidate-gene sequencing[1]. Recently, whole-exome sequencing combined with a filtering methodology was demonstrated as an approach to identify the gene underlying a mendelian disorder using a small number of affected individuals, with a proof-of-concept study that correctly identified the gene previously known to underlie Freeman-Sheldon syndrome[2]. Now, on page 30 of this issue, Michael Bamshad and colleagues[3] report the gene underlying an uncharacterized mendelian disorder, Miller syndrome, using the same strategy. Miller syndrome, also known as postaxial acrofacial dysostosis (MIM#263750), is a rare malformation syndrome that comprises anomalies including cleft palate, absent digits, ocular anomalies and others. The identification of the gene mutated in this disorder will allow improved diagnosis and a starting point for biological investigations, but the real advance of these two studies is the demonstration that this approach can be used to characterize the genetic basis of rare monogenic disorders.

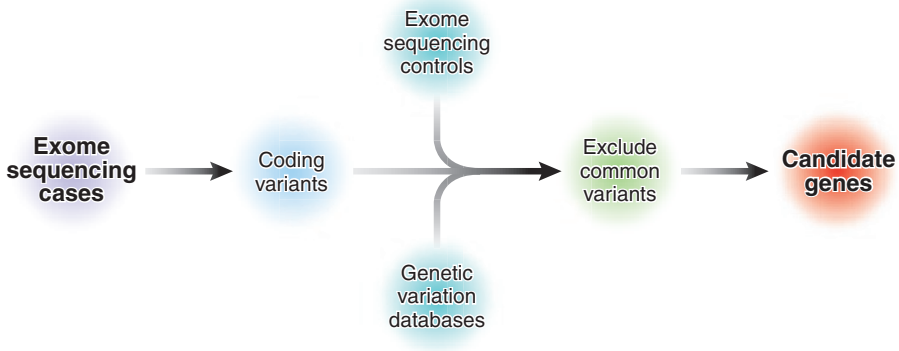

Figure 1 Exome sequencing and filtering strategy. In Ng *et al.*[3], the list of variants from the exome sequences of four individuals with Miller syndrome was first screened to select for genes found to have two nonsynonymous, splice site or indel sequence variants in each of the individuals. This list was then compared to the exome sequences of eight healthy controls[2] and dbSNP to exclude common variation and combined with a filtering strategy used to narrow the list of likely candidate genes underlying this rare disorder.

Exome sequencing approach

Ng *et al.*[3] sequenced the exomes of four individuals with Miller syndrome, including two siblings. The authors used an approach of targeted exome capture, including enrichment by array hybridization to 164,000 targets defining the exome, followed by sequencing in a massively parallel short-read sequencer to sequence at ~40-fold coverage. They used a stepwise filtering approach to screen the identified variants in order to select those likely to be implicated in the disorder (**Fig. 1**). They first screened for genes that contained nonsynonymous variants, splice site mutations or coding indels. They then compared the four exomes to those of eight control individuals (from HapMap and reported in ref. 2) and to the dbSNP database to exclude common variants. Finally, they excluded variants predicted not to be damaging by PolyPhen software. Although Miller syndrome was suspected to be recessive, they also tested a dominant model, and the recessive model fit the data best. They identified eight candidate genes under a dominant model, and only one, *DHODH*, under a recessive model. The four individuals with Miller syndrome were found to have six rare variants in *DHODH*. The excitement surrounding this study centers on the brute-force approach of exome sequencing combined with filtering that identified the disease-causing gene. This new approach will be critical to uncover the genes underlying rare mendelian traits, especially where the number of available individuals for study is small.

Ng *et al.*[3] interrogated four DNA samples with one technique (exome sequencing), but a set of exome data could also be interrogated with, for example, transcriptome sequences, proteomic data or methylation data. By integrating an exome dataset with a transcriptome dataset, one might identify abnormal mRNA isoforms caused by a previously unrecognized deep intronic splicing variant. Although Ng *et al.*[3] used a minimal amount of linkage data in their analysis (their sibling-pair requirement in the recessive model for the same two rare variants in the putative gene must be considered linkage analysis), one could envision using more linkage data in future analyses.

Leslie G. Biesecker is at the National Human Genome Research Institute, Bethesda, Maryland, USA.
e-mail: leslieb@helix.nih.gov

Although costs are not yet delineated, exome selection and sequencing is clearly less expensive than whole-genome sequencing at high coverage, making this approach more practical for groups studying rare mendelian traits.

New statistical metrics

The many ways in which genomic datasets could be combined in the search for genes underlying mendelian disorders suggests the need for new methods to assess statistical significance. Ng *et al.*[3] found six rare variants in the *DHODH* gene among four affected individuals and followed this by sequencing an additional three unrelated individuals with the disorder and a sibling in the second of the three initial families described above, for a total of 11 mutations found in six families. Together, this provides convincing evidence that the mutations in *DHODH* cause Miller syndrome. This also raises the question of how to assess significance, as there exists some threshold below which gene identifications by exome sequencing with filtering will arise by chance alone. Ng *et al.*[3] provide one approach by measuring the average frequency of 'new variants' per gene across the genome, as found by comparing their newly sequenced exomes to the common variation found in dbSNP. They squared that frequency to reflect the recessive model, cubed it for the three initial kindreds, and applied a Bonferroni correction for 17,000 genes. This yielded a significance threshold of 1.5×10^{-5}. Challenges to this approach will arise as it is generalized to other disorders. In addition, as the number of sequenced exomes rises, dbSNP will become populated by uncommon variants. We will need statistical metrics to distinguish false positives from true positives. We will also need metrics to account for locus and etiologic heterogeneity, which may often be unrecognized. Until these are available, we will have to rely on simple measures of coincidence and on supplemental proof such as animal models and functional studies.

Secondary findings

A fascinating finding was noted by Ng *et al.*[3]: the affected sibling pair here had a history of recurrent infections. It was difficult to determine whether this was an uncommon manifestation of Miller syndrome, a recessive contiguous gene syndrome or an unrelated disorder. Such complications continually bedevil clinicians who study and care for individuals with rare disorders, as most often the knowledge of the disorders and their molecular pathophysiology is insufficient to distinguish these possibilities. Ng *et al.*[3] found that these siblings were compound heterozygotes for *DNAH5*, a known cause of primary ciliary dyskinesia, which manifests as bacterial infections of the respiratory tract[4]. Therefore, the primary ciliary dyskinesia was coincidental to the Miller syndrome, has no implications for others with Miller syndrome and facilitates the care for the family under study.

This leads to the question of how many clinically relevant mutations reside in these four exomes or can be expected to be found in other sequencing projects. Should individuals participating in a sequencing study receive all or part of their sequence? How should results from minors be handled, especially those regarding adult-onset conditions or carrier status? Who would analyze these sequences, and what tools would they use? Who will deliver these datasets to the participants and interpret them and their clinical relevance? Is it appropriate to return results to patients in settings where it is impossible to implement care for the clinically relevant variants? To answer these questions, we will need clinical and behavioral studies of participants in genome sequencing studies, we will need their preferences and abilities to interpret these data, and we will need to explore different approaches to returning these data to participants and study how they use the data. The task going forward is to rapidly explore the multifaceted challenges associated with this technology so that we can not only discover the causes of rare diseases, but also move toward a future where whole-exome and eventually whole-genome sequences of individual patients lead to improvements in medical care.

1. Collins, F.S. *Nat. Genet.* **9**, 347–350 (1995).
2. Ng, S.B. *et al. Nature* **461**, 272–276 (2009).
3. Ng, S.B. *et al. Nat. Genet.* **42**, 30–35 (2010).
4. Olbrich, H. *et al. Nat. Genet.* **30**, 143–144 (2002).

Lung function and airway diseases

Scott T Weiss

Two studies report genome-wide association studies for lung function, using cross-sectional spirometric measurements in healthy individuals. They identify six genetic loci newly associated to natural variation in lung function, which may have implications for the related airway diseases of asthma and chronic obstructive pulmonary disease.

Asthma is a clinical syndrome defined by spontaneous or chemically induced increased airway responsiveness (bronchoconstriction) and reversible airflow obstruction (bronchodilation) with airway inflammation, often allergic in nature. In contrast, chronic obstructive pulmonary disease (COPD) is an airway disease with inflammation, but here the source of the inflammation is usually cigarette smoking, and it is defined by fixed, rather than reversible, airflow obstruction[1]. On pages 36 and 45 of this issue, two studies report genome-wide association studies (GWAS) identifying loci associated with lung function in predominantly healthy individuals as measured by spirometry[2,3]. Together, these studies report six loci newly associated with natural variation in lung function, bringing new insight into our understanding of the genetic basis of lung development and the related airway disorders asthma and COPD.

Asthma is primarily a disease of early childhood, with 80% of all cases diagnosed by 6 years of age. In contrast, COPD is a disease of later life, with most cases being diagnosed after age 60 years. Lung function abnormalities are present in both disorders. The primary environmental causes of both asthma and COPD are respiratory infections (viral and, in the case of COPD, bacterial) and cigarette smoking (passive and active).

Scott T. Weiss is at Harvard Medical School, Center for Genomic Medicine, and the Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts, USA.
e-mail: scott.weiss@channing.harvard.edu