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 tract⁴. 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 obstruction¹. 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

Although only 15% of active cigarette smokers develop COPD, 80% of individuals with COPD have increased airway responsiveness, one of the phenotypes defining asthma³. Thus, many investigators believe that smoking's effect on COPD risk is largely defined by the concomitant presence of asthma or increased airway responsiveness and by the ability of these intermediate phenotypes to reduce growth and accelerate decline in lung function.

Lung function measurements

Spirometry is defined by the dynamic maneuver of taking in a maximal deep breath and exhaling it as hard and as fast as one can. The forced expiratory volume in 1 second (FEV_1) is the amount of air that is blown out in 1 second after this maximal deep inhalation. The forced vital capacity (FVC) is defined as the volume of air expired after a maximal inhalation and then a maximal forced exhalation of 6 seconds duration or greater. The ratio of these two measures (FEV_1/FVC) is a measure of airflow obstruction. An FEV_1/FVC ratio less than 70% of that predicted after administration of a short-acting bronchodilator (beta-2 agonist) defines COPD. Repapi *et al.*² and Hancock *et al.*³ include measurements of both FEV_1 and FEV_1/FVC , but do not include measurements on the change in FEV_1 after the individual has taken either a bronchodilator (a measure of reversibility) or a bronchoconstrictor (a measure of airway responsiveness). These pharmacologically enhanced spirometric tests are measured as the change in FEV_1 at 15 minutes after two puffs of a short-acting beta agonist (bronchodilator) and the provocative dose of histamine or methacholine inducing a 20% decrease in FEV_1 from baseline (PD_{20} , increased airway responsiveness). Specific values of these tests (bronchodilator response >10% and PD_{20} <12.5 mg methacholine) define asthma and are also present in most individuals with COPD and some healthy individuals. These measures of lung function are correlated with each other and have been found to be heritable, with values of h^2 (the narrow-sense heritability due to genetic factors), as estimated from twin and family studies, ranging from 0.7 for FEV_1 and FVC to 0.35 for increased airway responsiveness—indicating that there is a substantial genetic component for all lung function measures^{4,5}.

Repapi *et al.*² performed a meta-analysis of 14 studies totaling 20,288 individuals as part of the SpiroMeta consortium, with replication in the Health 2000 study and the CHARGE Consortium study (21,209

Figure 1 Pulmonary function varies by age. The growth of FEV_1 as a percentage of maximal value shown is shown as a function of age. Curve A depicts reduced growth but normal decline. Curve B depicts premature decline in FEV_1 . Curve C depicts normal growth, no premature decline but accelerated decline. Curve D represents the optimal and ideal lung function, with normal growth and normal decline with age in a healthy individual. Individuals may show combinations of A–C, following a different curve at distinct life stages. Figure was adapted from reference 14.

individuals). Hancock *et al.*³ in the CHARGE Consortium performed a similar meta-analysis of 20,890 individuals with replication in the SpiroMeta consortium (20,228 individuals). These two studies each involved cross-sectional lung function data (FEV_1 and FEV_1/FVC measured at a single point in time), without use of a bronchodilator or a bronchoconstrictor. The study cohorts included predominantly individuals with normal lung function, although there were a small number of subjects with self-described asthma or COPD (the associations remained significant even after excluding individuals with these diseases)^{2,3}. They tested replication between these traits, performed an analysis stratified by ever- or never-smoking status, and also repeated analyses in a subset of the cohorts after excluding individuals with self-reported asthma and COPD. Finally, Repapi *et al.*² also performed gene expression studies on several of their candidate genes.

Genetics of lung function

The first GWAS report on lung function was reported by the Framingham Heart Study based on Affymetrix 100K data, suggesting an association to *GSTO2* on chromosome 10 to FEV_1 , although this has not been replicated⁶. Subsequently, a study reported a GWAS to the FEV_1/FVC ratio in 7,691 participants in the Framingham Heart Study, with replication in an additional 835 participants of the Family Heart Study⁷. They reported association of a locus on chromosome 4q31 with the percent of expected FEV_1/FVC ratio defined in a fashion identical

to the way it was defined in the two studies reported in this issue⁷. The associated variants were in an intergenic region upstream of *HHIP*, a hedgehog pathway gene with a known role in development. This locus was also reported to be associated with COPD, as defined by reduced FEV_1/FVC ratio, in an independent GWAS⁸. *SERPINE2* was selected from mouse lung-development microarray studies and then tested in a candidate gene association study with 127 probands from a family-based study of severe, early-onset COPD, with replication in 304 cases from the National Emphysema Treatment Trial and 441 controls from the Normative Aging Study⁹. Subsequent association studies confirmed this association for FEV_1/FVC in individuals with COPD¹⁰. Thus, two of the genes previously identified as associated to lung development have also been shown to be associated to COPD.

The first finding from the current studies was replication of the *HHIP* association to the FEV_1/FVC ratio^{2,3}. The two studies also report six new genetic loci associated with lung function and implicate a number of biologically plausible candidate genes. This includes associations at 6p21 near two genes with immune functions, *AGER* and *PPT2*; at 2q35 near *TSN1*, encoding an actin-binding protein; at 6q24.1 within *GPR126*, which encodes a G protein-coupled receptor; and at 15q33 within a thrombospondin family gene, *THSD4*. The associated locus at 5q32–33 includes *HTR4*, a serotonin pathway gene, and the chromosome 4q24 locus includes four genes including *GSTCD*. Five of the six

candidate genes (*AGER*, *TNS1*, *HTRA4*, *THSD4* and *GSTCD*) were shown to be expressed in lung tissue². These loci are all biologically plausible candidates that are involved in pathways known to be important in lung biology, including immune function, muscle function or inflammation. These studies did not replicate the *SERPINE2* association and also did not find evidence for association to the 15q *CHRNA5-CHRNA3-CHRNA4* locus previously associated with nicotine dependence, lung cancer and COPD^{11–13}.

Clinical relevance

Lung function levels depend strongly on age, height and gender, with distinct changes from childhood to adulthood (Fig. 1). Thus, because of study-design considerations, some gene variants identified in the two studies reported here may simply influence lung function in nondiseased subjects and will have no independent effect on patients with airways disease. However, the high within-

individual consistency of spirometric lung function measures over time, and the links of the identified candidate genes to lung development, suggest that genetic variations associated with lung function development might be important genetic determinants of lung function in both healthy individuals and those with airways disease (asthma and COPD). To test this hypothesis, the newly associated variants should be tested in both individuals with asthma and those with COPD. These studies should also be extended to include bronchodilator and bronchoconstrictor responses.

These loci account for a small proportion of the variation in FEV₁ and the FEV₁/FVC ratio. Repapi *et al.*² estimates that the five new loci identified in their study account for ~0.14% of the variation in FEV₁/FVC ratio. Translating this to clinical prediction remains further away, and the first steps toward this goal include additional validation, refinement of the genetic loci and determination

of the specific genes and their functional variants, as well as more accurate estimation of effect sizes from additional studies of unaffected and diseased subjects.

1. O'Connor, G., Sparrow, D. & Weiss, S.T. *Am. Rev. Respir. Dis.* **140**, 225–252 (1989).
2. Repapi, E. *et al. Nat. Genet.* **42**, 36–44 (2010).
3. Hancock, D.B. *et al. Nat. Genet.* **42**, 45–52 (2010).
4. Redline, S. *et al. Am. J. Epidemiol.* **129**, 827–836 (1989).
5. Xu, X. *et al. Am. J. Respir. Crit. Care Med.* **160**, 1928–1933 (1999).
6. Wilk, J.B. *et al. BMC Med. Genet.* **8** (Suppl. 1), S8 (2007).
7. Wilk, J.B. *et al. PLoS Genet.* **5**, e1000429 (2009).
8. Pillai, S.G. *et al. PLoS Genet.* **5**, e1000421 (2009).
9. DeMeo, D.L. *et al. Am. J. Hum. Genet.* **79**, 253–264 (2006).
10. Zhu, G. *et al. Am. J. Respir. Crit. Care Med.* **176**, 167–173 (2007).
11. Amos, C.I. *et al. Nat. Genet.* **40**, 616–622 (2008).
12. Hung, R.J. *et al. Nature.* **452**, 633–637 (2008).
13. Thorgeirsson, T.E. *et al. Nature.* **452**, 638–642 (2008).
14. Speizer, F.E. & Tager, I.B. *Epidemiol. Rev.* **1**, 124–142 (1979).

Getting connected in the globin interactome

Tobias Ragoczy & Mark Groudine

A new study provides compelling evidence that transcriptional regulation and three-dimensional genomic architecture are linked. The alpha- and beta-globin loci associate with hundreds of active genes across the genome at transcription factories in erythroid cells, and specialized Klf1-containing transcription factories mediate the association of Klf1-regulated genes.

The organization of a eukaryotic nucleus reflects its specific expression profile^{1,2}. On the genomic scale, this translates to preferred tissue-specific chromatin folding and chromosome positioning within interphase nuclei, including intermingling of looped segments within and between chromosomes^{3,4}. Consequently, co-regulated genes tend to localize near each other at activity hubs. However, the description of colocalized and coexpressed genes has been anecdotal and limited to a few genes. What had been missing until recently^{5,6} was information on how extensive and consistent these associations are at the genomic level. On page 53 of this issue, Peter Fraser and colleagues⁷ use an array of biochemical and cytogenetic approaches to elegantly describe

such a genomic transcription interactome by exploring the gene associations of the mouse α - (*Hba*) and β -globin (*Hbb*) loci in erythroid cells.

Globins cast a wide net

The globin genes are highly expressed in developing erythrocytes, and they consistently associate with transcription factories, which are foci of hyperphosphorylated RNA polymerase II (RNAPII) and are sites of active transcription. Using double-label RNA fluorescence *in situ* hybridization (FISH), Schoenfelder *et al.*⁷ discovered that several expressed erythroid-specific genes (of a set of 33 examined loci on 15 chromosomes) frequently localize to the same transcription factories as *Hba* or *Hbb*. To expand on this initial observation, Schoenfelder *et al.* used a modified chromosome conformation capture technique (enhanced ChIP-4C, or e4C) to establish the genomic interactome of the globin loci. They precede the assay with a chromatin immunoprecipitation (ChIP)

against hyperphosphorylated RNAPII and a bait-specific (globin) enrichment step involving a biotin pulldown. These clever adaptations ensure that the detected associations are restricted to actively transcribed genes and thus result in a highly improved signal-to-noise ratio. Consequently, the results provide the first truly robust and reproducible large-scale detection of long-range and interchromosomal associations between active genes. Microarray analysis of the e4C material confirms known interactions between *Hbb* and erythroid-specific genes (for example, *Eraf* and *Uros*) and hundreds more. Strikingly, around 90% of these detected interactions were in *trans* for both *Hbb* and *Hba*. The authors performed elaborate controls to demonstrate that virtually all ligation products obtained via the e4C protocol originated from individual cross-linked complexes and thus represent bona fide long-distance interactions. Moreover, paired-end tag sequencing (PET) of phosphorylated RNAPII ChIP material confirmed

Tobias Ragoczy and Mark Groudine are in the Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
e-mail: markg@fhcrc.org