The Next Generation of Complex Lung Genetic Studies

Ivana V. Yang¹ and David A. Schwartz¹

¹Department of Medicine, University of Colorado Denver, Aurora, Colorado

Common genetic risk variants identified by genome-wide association studies have explained a small portion of disease heritability in complex diseases. It is becoming apparent that each gene/locus is heterogeneous and that multiple rare independent risk alleles across the population contribute to disease risk. Next-generation sequencing technologies have reached the maturity and low cost necessary to perform whole genome, whole exome, and targeted region sequencing to identify all rare risk alleles across a population, a task that is not possible to achieve by genotyping. Design of whole genome, whole exome, and targeted sequencing projects to identify disease variants for complex lung diseases requires four main steps: library preparation, sequencing, sequence data analysis, and statistical analysis. Although data analysis approaches are still evolving, a number of published studies have successfully identified rare variants associated with complex disease. Despite many challenges that lie ahead in applying these technologies to lung disease, rare variants are likely to be a critical piece of the puzzle that needs to be solved to understand the genetic basis of complex lung disease and to use this information to develop better therapies.

Keywords: genetic variants; rare variants; next-generation sequencing; complex lung disease; disease risk alleles

Complex disease genetics has focused on the search for common genetic risk variants (allele frequencies > 5%) that influence common diseases and phenotypes. The basis for this search is the common disease common variant (CDCV) hypothesis, which assumes that a relatively small number of ancient common risk alleles exist that each confer small to moderate risks of disease. Under this model several weak common genetic (and environmental) risk factors jointly result in the development of disease. Several practical considerations have made the CDCV quite amenable to exploration in disease genetics, including (1) the high frequency of putative risk alleles, which allows their screening by low-cost genotyping; (2) the extensive linkage disequilibrium (LD, correlation) exhibited between common variants in the human genome, which allows the screening of many more markers indirectly by LD mapping; and (3) the development of high-throughput genotyping arrays, which have allowed genome-wide association (GWA) studies. As a consequence, more than 1,500 GWA studies exploring the role of common variants in several hundred complex diseases or phenotypes have been

conducted, resulting in the identification of a great number of common disease variants (<http://www.genome.gov/gwastudies/>).

However, it has become apparent that a significant amount of heritability remains unexplained for most complex common diseases even after intelligently designed GWA studies screening more than 1 million markers in large groups of case and control subjects. The “missing heritability” problem resultant from these GWA studies has triggered interest in the potential role of rare variants in complex disease, as indicated by the common disease rare variant (CDRV) hypothesis. Under this hypothesis, any given risk gene or locus is characterized by high allelic heterogeneity, namely, these risk loci contain multiple rare independent risk alleles across the population, each with moderate to high penetrances. Direct sequencing rather than genotyping is required for exploration of the CDRV hypothesis in complex diseases, as a result of the expected allelic heterogeneity, to identify all rare risk alleles across a population. Second, screening of these rare alleles is not amenable to LD tagging approaches as rare alleles are poorly tagged by common variants and individual rare variants are expected to occur on different haplotypes (Figure 1). The emergence of massively parallel sequencing technologies has dramatically reduced the time and cost of study population sequencing, setting the stage for widespread exploration of the CDRV hypothesis at both the gene or locus and genome or exome scale, similar to the growth in CDCV studies. Table 1 provides definitions for terms commonly used in the remainder of this article.

SEQUENCING TECHNOLOGIES

The first published sequences of the human genome (1, 2) were accomplished by automated Sanger sequencing with dideoxy chain termination (3). This “first-generation” sequencing technology is capable of generating up to an 800-bp sequence and 96 samples per run, and it has been estimated that the cost was approximately \$2.7 billion to produce a draft sequence of the human genome (<http://www.genome.gov/11006943>). Although this technology is still widely used to sequence individual genes and loci, genetic analysis of the human genome now relies on the “second generation” of the next-generation sequencing (NGS) technologies. NGS refers to a group of strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods, with each run generating gigabases (Gb) of sequence data. Second-generation technologies (summarized in Table 2) require clonal amplification of sequencing templates, are manufactured by three commercial vendors, and are currently broadly in use (4). Using this new technology, the cost of sequencing the human genome at 30-fold coverage is \$5,000 to \$7,000, depending on the platform.

In addition, there are several single-molecule sequencers under development, most notably from Pacific Biosciences (Menlo Park, CA) and Helicos BioSciences (Cambridge, MA). These technologies differ in that they do not require the clonal amplification of molecules to be sequenced, but rather a single DNA molecule is sequenced by synthesis using a DNA polymerase.

(Received in original form June 8, 2012; accepted in final form August 15, 2012)

Supported by NIH grants R01-HL095393, RC2-HL101715, and R01-HL097163, and by VA Merit Award 1101BX001534.

Correspondence and requests for reprints should be addressed to Ivana V. Yang, Ph.D., University of Colorado Denver, 12700 East 19th Avenue, 8611, Aurora, CO 80045. E-mail: ivana.yang@ucdenver.edu

Am J Respir Crit Care Med Vol 186, Iss. 11, pp 1087–1094, Dec 1, 2012

Copyright © 2012 by the American Thoracic Society

Originally Published in Press as DOI: 10.1164/rccm.201207-1178PP on August 30, 2012

Internet address: www.atsjournals.org

Figure 1. Schematic illustrating the relationship of rare and common variants to linkage disequilibrium (LD) blocks in the genome. Linkage disequilibrium refers to the nonrandom associations of alleles at different loci. Variants 1–5, 8, and 9 are single-nucleotide polymorphisms (SNPs; >5% population frequency) whereas variants 6 and 7 are rare variants (<1% population frequency). SNPs 1–5 in block 1 are not independent and therefore in LD with each other, whereas they are not in LD with SNPs in block 2 (SNPs 8 and 9). Only one of the five SNPs in block 1 need be genotyped to capture the variation in block 1. Similarly, only one of two SNPs in block 2 must be genotyped to capture the variation in block 2. In addition to occurring infrequently, rare variants are not in LD with any common polymorphisms and therefore must be sequenced. Values in the LD map represent pair-wise $100 \times D'$ values of linkage disequilibrium. The *blank squares* represent D' values of 1.0 (complete LD). Strong LD is indicated by *red squares*, whereas *pink squares* and *white squares* indicate uninformative and low confidence values, respectively.

Single-molecule sequencing technologies promise much simpler sample preparation with fewer biases and dramatically lower input sample requirements in the future. This technology is reviewed elsewhere (5).

DESIGN OF STUDIES TO IDENTIFY RARE GENETIC RISK VARIANTS FOR COMMON LUNG DISEASES

Design of whole genome, whole exome, and targeted sequencing projects to identify disease variants for complex lung diseases requires four main steps: library preparation, sequencing, sequence data analysis, and statistical analysis. Presumably most of these studies will be focused on the identification of rare genetic variants. Whole genome sequencing (WGS) studies offer the advantage of capturing variation in noncoding variants and identifying structural variants (SVs). Whole exome sequencing (WES) studies are focused on coding variants (although newer target enrichment strategies capture untranslated regions and some promoter and intronic sequence); are about 10-fold less expensive than WGS studies, thus allowing a larger number of samples to be profiled; and the data produced are less complex to analyze. Targeted resequencing studies are well suited for following up on GWA studies or linkage loci and are generally designed to capture the entire locus, including both coding and noncoding regions. Loci with previously associated common variants are more likely to contain functional rare variants and therefore may be the most fruitful application of next-generation sequencing to complex disease. It was initially proposed that two affordable strategies for identification of disease-causing variants were (1) sequencing of affected individuals in a pedigree followed by genotyping of candidate variants to demonstrate cosegregation with disease in the family and (2) extreme-trait sequencing of a small number

of individuals at the tails of the trait distribution followed by targeted sequencing or genotyping in a larger cohort (6). As the cost of sequencing has decreased, other study designs have been adopted. As sequence data have accumulated, it is becoming clear that the number of identified rare variants is growing much more rapidly than predicted by the neutral model with constant population size, and that this is at least in part due to purifying selection (7). As a consequence, rare variants are enriched for those with deleterious effects. Investigators are taking advantage of this fact and selecting loci/genes for targeted sequencing based on the strength of evidence for purifying selection (8).

Library Preparation

Despite differences in enzymology, chemistry, high-resolution optics, and hardware and software used, all second-generation sequencing technologies encompass the same series of steps in data collection (Figure 2A). With the exception of WGS applications, the first step in the process is target enrichment to capture a DNA sequence of interest. Several approaches based on long-range polymerase chain reaction (PCR), microdroplet PCR, in-solution hybridization, and array hybridization are commercially available. The choice of the target enrichment strategy depends on the length of target DNA and the number of samples to be sequenced in a given project (discussed in Reference 9). The three most commonly used exome capture technologies have been compared, highlighting the advantages and disadvantages of each particular platform (10). After or during target enrichment, platform-specific adaptors are ligated to sheared DNA to create DNA libraries. In this same step, barcode sequences can be introduced to allow for multiplexing of several samples in one sequence run. Sequencing of DNA pools rather than barcoding and sequencing of individual samples has been used as a cost-efficient strategy for variant identification in a larger number of individuals (11). In this approach, next-generation sequencing of DNA pools of 10 or more individuals allows for identification of a small number of pools with individuals harboring the rare variants. This is followed by capillary sequencing or genotyping of individuals to identify individuals harboring the variants. Although the pooling approach has been successfully applied (12), it has also been shown to have limitations (13). With the high level of sample barcoding and multiplexing available on all platforms and sequencing costs rapidly decreasing, sequencing of a large number of individuals is increasingly becoming more realistic. However, the labor and cost of target enrichment and library preparation for a large number of individuals are still substantial.

Sequencing

Adaptors enable attachment of DNA templates to beads for emulsion PCR (SOLiD [Life Technologies, Carlsbad, CA] and 454 [(Roche, Basel, Switzerland)] or to the oligonucleotide-derivatized surface of a flow cell for solid-phase amplification (Illumina, San Diego, CA) to carry out clonal amplification. To perform sequencing and image analysis, beads are attached to a chemically modified glass surface (SOLiD) or placed in the wells of a picotiter plate (454), whereas no additional step is required for the Illumina technology. Finally, sequencing and imaging are accomplished by platform-dependent sequencing chemistry and imaging technology. The amount of sequence and the number of sequencing runs required depend on the desired depth of coverage. Early whole genome sequencing studies were performed at 30- to 40-fold coverage whereas higher coverage (100- to 200-fold) has been used in more recent whole exome and targeted resequencing studies.

TABLE 1. DEFINITIONS OF TERMS COMMONLY USED IN NEXT-GENERATION SEQUENCING RESEARCH

Term	Definition
Adaptor	A short DNA sequence that is ligated onto the ends of DNA fragments to provide priming sequences for both amplification and sequencing; adaptors are platform specific
Burrows–Wheeler transform (BWT)	The foundation of algorithms for compression and indexing of text data
Clonal amplification	PCR on a single DNA molecule to create copies of the DNA molecule
Copy number variant (CNV)	A form of structural variation; large segments of DNA, ranging in size from thousands to millions of DNA bases, that vary in copy number
Depth of coverage	Number of sequence reads at any particular base, averaged over the entire sequence target; calculated using the Lander–Waterman equation $C = LN/G$ where L = read length, G is the length of the target sequence (exome, genome, targeted area), and N is the number of reads
Error rate	Proportion or percentage of miscalled bases in the sequence; occurs as a result of errors in base calling or alignment
Evolutionary constraint	Lack of change in sequence during evolution; sequence conservation is used as an indicator of regulatory potential
Hashing	A procedure used to accelerate alignment by storing information about where in the reference genome a particular substring or subsequence occurs
Linkage disequilibrium	The nonrandom association between two or more alleles such that certain combinations of alleles are more likely to occur together than other combinations of alleles
Mapped reads	Number (or percentage) of reads mapped to the reference genome; a read is uniquely mapped if its second-best hit contains more mismatches than its best hit
Penetrance	Proportion of individuals carrying a variant that also express an associated trait
Population structure	Individuals from a certain geographic area are more closely related to one another than those randomly selected from the general population
Purifying (or negative) selection	Selective removal of alleles that are deleterious
Quality values (QVs)	Phred quality score; $Q_{\text{phred}} = -10 \log_{10} P(\text{error})$; 20 corresponds to a 1% error in base calling and is the most commonly used cutoff
Sample barcoding	A short DNA sequence that is ligated onto the ends of DNA fragments preceding adaptors to allow for multiplexing of samples in one sequence run
Shotgun library	Library consisting of DNA sheared randomly into fragments; fragmentation is generally accomplished with sophisticated sonicators that allow for precise shearing into the desired size range
Single-nucleotide polymorphism (SNP)	Variant with a population minor allele frequency (MAF) > 5%
Single-nucleotide variant (SNV)	Variation at a single base position in the genome
Structural variant (SV)	Variation in structure of the chromosome; comprises insertions, deletions, rearrangements, and CNVs
Target enrichment	A set of strategies for enrichment of region(s) of the genome for sequencing; choice of specific strategy depends greatly on the size of the region of interest

Sequence Data Analysis

Data analysis workflow begins with the base calling and alignment of sequence data to the reference genome (Figure 2B) (14). Base-calling procedures vary from platform to platform and are prone to different types of errors (15). These programs generate per-base quality values (QVs) that are typically converted to Phred-like quality scores. Most alignment algorithms (see Reference 15 for a list of noncommercial aligning software) are based on either *hashing*, a procedure for creating a data structure that helps to accelerate alignment, or an effective data compression algorithm (Burrows–Wheeler transform; BWT). BWT-based aligners are fast, memory efficient, and especially useful for aligning repetitive reads; however, they are less sensitive than the state-of-the-art hash-based algorithms. After alignment, Phred-like quality scores can be recalibrated to account for any errors due to misalignments. Most software programs provide a number of additional run metrics that allow the user to assess quality of sequence data; these include number of raw reads, number of mapped reads, number of unique reads, sequence coverage, alignment scores, and error rates. The quality control metrics allow researchers to determine potential experimental and alignment biases and to remove low-quality and poorly mapping reads from further analysis. The high-quality, aligned reads are then analyzed to identify DNA sequence variants, including single-nucleotide variants (SNVs), structural variants (SVs), and copy number variants (CNVs), in each sample of interest compared with the reference sequence (Figure 2C). Multiple commercial and free software options also exist for variant calling (noncommercial packages are listed in Reference 15). SNV calling algorithms are divided into two main classes: (1) those that perform variant calls in individual samples (such as SOAP2)

and (2) those that use multiple samples to call variants (such as SAMtools). Prior studies have found about 4 million SNVs in WGS studies and more than 20,000 SNVs in an exome in comparison with a reference sequence, although these numbers will vary depending on the depth of coverage across the area sequenced (16). Although these packages use filtering steps to exclude low-quality variants, further quality-based filtering of the data was done after variant calling in early sequencing studies to reduce the number of false positives. More recent approaches for SNV quality control in complex traits rely on population genetic statistics and properties of human genetic variation (implemented in packages such as IMPUTE2) (7).

As an alternative to multistep filtering, studies have also employed more formal statistical approaches. One of these methods provides fast computation of approximate P values for individual genes, adjusts for the background variation in each gene, allows for incorporation of functional or linkage-based information, and accommodates designs based on both affected relative pairs and unrelated affected individuals (17). A different unified framework for variant discovery that does not involve a formal statistical approach but consists of three steps: (1) data processing, (2) variant discovery, and (3) integration with known variants and other information, such as pedigrees and population structure to recalibrate variant quality, has also been developed by another group (18). Both strategies were applied to published data sets and resulted in high-quality variant discovery. Several other pipelines that incorporate linkage disequilibrium (LD) information have also been developed (15). These and other sophisticated approaches for analysis of next-generation sequence data are likely to replace multistep filtering approaches in the near future.

TABLE 2. SECOND-GENERATION SEQUENCING PLATFORMS PRESENTLY IN WIDESPREAD USE IN GENETIC STUDIES OF COMPLEX DISEASES

Platform	Read Length	Common Use	Library Types	Clonal Amplification Strategy	Sequencing Strategy	Advantages	Disadvantages
Solexa (Illumina, San Diego, CA)	35–100 bp	WGS, WES	Fragment, paired-end	Solid-phase PCR	Sequencing by synthesis uses a reversible terminator-based method that enables detection of single bases as they are incorporated into growing DNA strands. A fluorescently labeled terminator is imaged as each dNTP is added and then cleaved to allow incorporation of the next base. Incorporation bias is minimized by the natural competition between the reversible terminator-bound dNTPs present during each sequencing cycle.	Low reagent cost, high accuracy	Difficulties with alignment of short reads, particularly in repetitive regions of the genome
SOLiD (Life Technologies, Carlsbad, CA)	35–100 bp	WGS, WES	Fragment, mate-pair, paired-end	Emulsion PCR	Sequencing by ligation, in which case a set of four fluorescently labeled dibase probes competes to ligate to the sequencing primer, with the specificity of the dibase probe achieved by interrogating every first and second base in each ligation reaction. Multiple cycles of ligation, detection, and cleavage are performed, with the number of cycles determining the eventual read length.	Low reagent cost, high accuracy	Difficulties with alignment of short reads, particularly in repetitive regions of the genome
454 (Roche, Basel, Switzerland)	750 bp	Targeted resequencing	Fragment, paired-end	Emulsion PCR	Pyrosequencing, which allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step through chemiluminescence.	Easy alignment of long reads	Higher reagent cost, lower accuracy
Ion Torrent (Life Technologies)	400 bp	Targeted resequencing	Fragment and mate-pair	Emulsion PCR	Post-Light sequencing technology; the first platform to avoid the cost and complexity associated with four-color optical detection. PGM makes base calls by detecting protons released from the incorporation of unmodified dNTPs by a natural DNA polymerase. The protons are detected as a pH change in wells of a semiconductor chip.	Adaptable run sizes (10-Mb, 100-Mb, and 1-Gb chips)	Difficulties in sequencing homopolymers

Definition of abbreviations: bp = base pairs; dNTP = deoxyribonucleoside triphosphate; PCR = polymerase chain reaction; PGM = Personal Genome Machine; WES = whole exome sequencing; WGS = whole genome sequencing.

Association Testing

Candidate gene/locus-based rare variant studies may be particularly fruitful given the large number of risk loci identified by both GWA and linkage studies. In considering a gene-based model of disease risk, it is possible that loci harboring common risk alleles may also contain rare risk alleles. Second, the failure to map causative common risk alleles to GWA loci has spurred the hypothesis that these GWA signals are actually marking multiple rare risk alleles, producing so-called “synthetic associations.” Although case-control association studies similar to common variants studies can be employed to explore rare variants, the experimental and analytical design will vary, based on many factors including the rarity of alleles and the sample size tested. In general, it is believed that causative rare variants will range in frequency from 0.1 to 1.0%. Despite the tendency to concentrate sequencing only on case subjects, to enrich for rare causal variant discovery, followed by genotyping in control subjects, this strategy has been shown to result in excess type I error in testing rare variants for association. Rather, a balanced number of case and control subjects should be sequenced to control for type I error. Identified variants could be tested by single-marker, multimarker, or collapsing methodologies (19). Despite the expectation of high effect sizes for rare risk alleles (relative risks, ~2.5–5.0) and large sample sizes adequately powered for GWA studies, most study populations will be underpowered to conduct univariate single-marker tests of association (χ^2 or Fisher’s exact test) because of the low frequency of the rare alleles tested and the necessity to correct for multiple testing. Alternatively, the information from multiple variant sites can be combined in a single multivariate test (Hotelling T^2 , multivariate regression) to avoid the multiple testing problems of univariate tests. However, the power of these tests suffers because of the increase in test degrees of freedom and power can be further adversely affected by the inclusion of nonrisk rare alleles in the testing variable. Another alternative is the collapsing of rare variant genotypes across the risk locus or gene tested into a single dichotomous variable that indicates the presence of at least one or absence of a rare risk variant. The general approach is to test for association between disease status and the accumulation of rare variants across the risk locus or gene units rather than with any single variant. Three main classes of collapsing tests are (1) tests that apply group summary statistics on variant frequencies in case and control subjects (2); those that test for similarity in unique DNA sequences in different individuals; and (3) regression models that test collapsed sets of variants (and other variables) as predictors of the phenotype. One of the commonly used tests from the first class of tests is the combined multivariate and collapsing (CMC) method (20), which has the advantage of appropriately controlling type I errors even when nonfunctional variants are included in the test. Extensions of the CMC test include the aggregated number of rare variants (ANRV) method (21), which jointly assesses the role of rare and common variants and incorporation of the CMC statistic in regression model tests. A comprehensive list of tests for association of rare variants with complex disease is included in Reference 22.

After association testing, identified variants can be segregated into groups of previously discovered variants, for example, ones present in the 1000 Genomes Project database or Single Nucleotide Polymorphism Database (dbSNP), or novel variants. More importantly, they are analyzed for functional consequences (e.g., coding or noncoding, nonsynonymous or synonymous, promoter, evolutionarily conserved) to try and discern their possible relevance to the traits studied. The approaches to filter the large number of benign variants from the few risk variants is evolving almost

as rapidly as sequencing technologies; more recent advances and challenges in this field are reviewed in detail in Reference 23. Two general classes of algorithms that predict deleteriousness of protein-coding variants are (1) those that explicitly define the evolutionary property of deleterious variants and make predictions on the basis of similarity to that definition (first-principle approaches) and (2) trained classifiers, which use heuristic combinations of a number of properties that distinguish a set of true positives from negatives to generate prediction rules (comprehensive list available in Reference 23). Nonsense and frameshift mutations result in loss of protein function and are therefore prioritized as the strongest candidates followed by nonsynonymous changes in the coding sequence. SIFT (a first-principle approach) (24) and PolyPhen (a trained classifier) (25) are the two most commonly used algorithms to predict the consequence of amino acid substitution variants on protein function. In fact, SIFT has been incorporated into the ANNOVAR pipeline (26) whereas PolyPhen is a part of the SeattleSeq pipeline (<http://snpgs.washington.edu/SeattleSeqAnnotation>) as well as the PLINK/SEQ suite (<http://atgu.mgh.harvard.edu/plinkseq/>). ANNOVAR and SeattleSeq are tools that provide comprehensive functional annotation of variants, both known and novel; the annotation includes database IDs, gene names, protein positions and amino acid changes, conservation scores, HapMap frequencies, and predictions for deleteriousness. Methods for prioritizing noncoding variants based on nucleotide sequence conservation are also being used in large-scale sequencing studies; of the many available algorithms (listed in Reference 23), one commonly used is the genomic evolutionary rate profiling (GERP) algorithm (27). These algorithms use comparative genomics, generally limited to mammalian species as nucleotide sequence is less conserved than protein sequence, to estimate nucleotide-level evolutionary constraints in genomic sequence alignments and to assign conservation scores. Higher conservation scores are indicative of a more likely regulatory function and can be used to prioritize noncoding variants for further studies.

An alternative to post hoc analysis of variants in associated genes/loci is to incorporate functional information into the test and to stratify or weight rare alleles by functional significance. A number of tests allow for inclusion of prediction scores in test statistics; PLINK/SEQ, for example, includes previously computed PolyPhen scores (7). The reasoning behind incorporating functional information is to avoid the possibility of most individuals possessing variants in a specific region, based on simple counting of individuals with variants. Another way to avoid this issue is to weigh by frequency of variants. An additional advantage of incorporating functional information in the statistical test is that the analysis is more likely to identify disease-predisposing variants (22).

As in all genomic studies, one of the challenges in the analysis of sequence data concerns adjustment for multiple comparisons. The traditional Bonferroni correction to correct for tests in all genes is generally too stringent and one of the commonly used methods relies on the *i*-stat from PLINK/SEQ to set a threshold and correct only for the number of genes for which there is power to detect association (7). Another commonly used strategy is to create an empirical distribution of *P* values by permuting phenotypic data and comparing the minimal *P* value from the real data to that distribution (7).

VALIDATION

The first step in the validation process is confirmation (or internal validation) of the presence of rare variants by an independent technology. The most cost-effective way to validate variants with low frequencies is to directly sequence the region that contains

Figure 2. Overview of sequencing and data analysis strategies applied to whole genome, whole exome, or targeted region data to identify rare variants associated with complex traits. (A) In “second-generation” sequencing technologies, DNA is sheared and platform-specific adapters are ligated to allow for shotgun library construction followed by sequence capture (omitted in WGS), clonal amplification to obtain enough material, and sequencing of clonally amplified fragments. (B) Sequence reads from (A) are aligned with the reference sequence to obtain a consensus sequence for each sample. (C) The consensus sequence for each sample from (B) is used for variant calling in each sample compared with the reference. Some variant-calling algorithms perform this step on individual samples whereas others perform multisample variant calls. This is followed by quality filtering to remove low-quality calls. High-quality calls are then used in association tests that collapse rare variants in a gene or locus. Gray screen-highlighted case/control sequence indicates an associated gene/locus due to statistical enrichment of rare variants across case subjects compared with control subjects in that region. Finally, variants are characterized on the basis of the current knowledge of genetic variation into known and potentially novel variants, and the consequences of the variant on protein function and evolutionary conservation. 1000 Genomes = 1000 Genomes Project database; dbSNP = Single Nucleotide Polymorphism Database; HapMap = International HapMap (haplotype map) Project database; WGS = whole genome sequencing.

multiple rare variants in different individuals; this is most commonly achieved by Sanger sequencing. As an alternative to Sanger sequencing, an independent next-generation platform with high coverage can be used. For example, WES variants identified on the Illumina or SOLiD platform can be easily sequenced on the Ion Torrent platform for confirmation. On the other hand, variants with higher frequencies can be genotyped at a lower cost than sequencing.

The second and much more complex step is replication (or external validation) of results in an independent cohort. The choice of the technology for the external validation is straightforward, as the same assays used for internal validations can be used. The complexities lie in the selection of the independent cohort and depend greatly on the study design (family-based, case-control, extremes of phenotypes). Most complex lung diseases are heterogeneous in their clinical presentation, the underlying biology, as well as population demographics such as race; these factors need to be taken into account in the design of replication studies.

HUMAN GENETIC STUDIES USING NEXT-GENERATION SEQUENCING

Just since 2009, a number of studies have used next-generation sequencing to characterize normal human genetic variation as well as variation associated with disease development. In the area of normal human variation, the 1000 Genomes Project, which aims to identify most genetic variants that have frequencies of at least 1% in the populations studied, completed the pilot phase of the

project, which consisted of sequencing the complete genomes of 179 individuals from West Africa, Europe, China, and Japan at low coverage; sequencing the complete genomes of two sets of trios (a child and both its parents) at high coverage; and sequencing the exome regions for an additional 697 individuals. The results of the pilot phase identified 15 million SNPs, 1 million short insertions and deletions, and 20,000 SVs, most of which were previously undescribed (28, 29). This and another WGS study of a family quartet (30) estimated the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. These studies are providing a foundation for investigating the relationship between genotype and phenotype.

Several early studies used WGS or WES to identify rare variants in mendelian disorders (31–33), and WES continues to prove useful for identification of causal genes in mendelian disorders (34). Identification of causal variants in complex disease is certainly much more challenging, but studies are proving its feasibility when clever experimental design is used. For example, WES of 20 individuals with sporadic autism spectrum disorder (ASD) and their parents (family trios) resulted in identification of a number of novel candidate genes for ASD (35). In another study, WES on two distant members of a large family with autosomal dominant inheritance of thoracic aortic aneurysm leading to acute aortic dissections (TAAD) identified mutations in *SMAD3*. Sanger sequencing of 181 probands with familial TAAD identified 3 additional *SMAD3* mutations in 4 families, resulting in a combined logarithm of odds score of 5.21 and mutations in *SMAD3* responsible for 2% of familial TAAD (36).

However, next-generation sequencing technologies are likely to have the highest impact in the area of complex disease when used to sequence candidate regions and genes from GWA studies and linkage studies to identify rare variants (37, 38). To follow up on GWA study results, 144 target regions that covered exons and regulatory sequences of 10 type 1 diabetes (T1D) candidate genes (total, 31 kb) were resequenced in pools of DNA from 480 patients and 480 control subjects and then tested for association with disease in more than 30,000 individual subjects (37). This analysis identified four rare variants in *IFIH1* that are predicted to alter the expression and structure of IFIH1 (MDA5), a cytoplasmic helicase that mediates induction of the interferon response to viral RNA. This study found that rare alleles of all associated *IFIH1* polymorphisms consistently protect from T1D, independent of each other, whereas *IFIH1* alleles carried by the majority of the population predispose to the disease, suggesting that variants that disrupt IFIH1 function in the host antiviral response have been negatively selected. In another study, GWA was performed to identify seven loci associated with hypertriglyceridemia (HTG), followed by resequencing of protein-coding regions of four GWA study candidate genes in an independent cohort (38). Sequence analysis identified a significant burden of 154 rare missense or nonsense variants in 438 patients with HTG, in contrast to 53 variants in 327 control subjects ($P = 6.2 \times 10^{-8}$), corresponding to a carrier frequency of 28.1% of patients with HTG and 15.3% of control subjects. This study also suggested that rare variants found in four GWA study-identified genes incrementally contribute to the unexplained variation underlying HTG pathophysiology, based on a logistic regression model including clinical variables and both common and rare genetic variants. This model explained 41.6% of total variation in HTG diagnosis: clinical variables explained 19.7%, common genetic variants in seven HTG-associated loci explained 20.8%, and rare genetic variants in four HTG-associated genes explained 1.1%. Finally, the first study of rare variants in lung disease used traditional Sanger sequencing to sequence coding exons and flanking noncoding regions of 9 genes that showed the strongest signatures of purifying selection among 53 candidate asthma-associated genes (8). Sequencing of 450 (108 European American and 342 African American) case subjects and 515 (248 European American and 267 African American) control subjects identified rare variants in four genes (*AGT*, *DPP10*, *IKBKAP*, and *IL12RB1*) associated with asthma susceptibility among African Americans. Rare variants in *IL12RB1* also contributed to asthma susceptibility among European Americans, but the majority of rare variants in this genes were population specific. Overall, the contribution of rare variants to asthma susceptibility was due predominantly to noncoding variants in sequences flanking the exons. However, nonsynonymous rare variants in *DPP10* and in *IL12RB1* were associated with asthma in African Americans and European Americans, respectively. These studies illustrate the promise of sequencing of GWA study loci to aid in identifying candidate genes on the basis of the presence of rare variants that contribute to the missing heritability in GWA studies.

PROMISE OF SEQUENCING TECHNOLOGIES FOR GENETIC ANALYSIS OF LUNG DISEASE

At present, the application of high-throughput sequencing technologies to lung disease has been limited to only a few publications. Two publications performed whole-genome sequences of lung malignancies. Pleasance and colleagues compared a small-cell lung cancer cell line derived from a bone marrow metastasis with an Epstein-Barr virus-transformed lymphoblastoid line from the same individual (39), whereas Lee and colleagues sequenced DNA isolated from an adenocarcinoma and adjacent normal tissue from the same individual (40); both studies identified extensive somatic mutations associated with lung cancer and tobacco smoke exposure.

Although the promise of next-generation sequencing technologies in understanding the genetic basis of complex lung diseases has been largely unrealized to date, the foundation for this work has been established. The 1000 Genomes Project is a general resource for human genetic variation and provides reference for different ethnic populations; using sequence variation reported for the race/ethnicity of the study population is essential in reducing numbers of false positive associations with disease development. The National Heart, Lung, and Blood Institute-sponsored Exome Sequencing Project (ESP) is sequencing protein-coding regions of the human genome across diverse, richly phenotyped populations with heart, lung, and blood disorders and has created the exome variant server (<http://evs.gs.washington.edu/EVS/>) for the community to be able to use the data. Their first publication identified *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis (41).

The promising area for application of next-generation sequencing technologies is to identify, through GWA or linkage studies, rare variants in genes and loci in which common polymorphisms have previously been associated with lung disease. These include a number of published GWA studies of asthma, chronic obstructive pulmonary disease, and pulmonary function (reviewed in Reference 42) and our linkage study in idiopathic pulmonary fibrosis (43), among others. Analogous to published sequencing studies of GWA study loci from other diseases, we believe that sequencing of loci associated with lung disease will unravel rare variants whose accumulation will explain a portion of the “missing heritability” in the genetics of complex lung disease. However, identifying rare variants in addition to common polymorphisms may not be enough to identify disease-causing genes. Integration of genetic data with gene expression and epigenetic data that are also being collected on next-generation sequencing platforms will likely be necessary to prioritize candidates for further studies (44). One example of such integrative analysis is expression quantitative trait loci (eQTL) mapping of common and rare variants (45); however, this could be extended to methyl-QTLs. Once prioritized, candidate genes need to be studied for their function in disease pathogenesis, using cellular and animal studies. Despite many challenges that lie ahead in terms of sequence data collection, analysis, and integration, genetic variants identified by sequencing technologies are likely to be a critical piece of the puzzle that needs to be solved to understand the genetic basis of complex lung disease and to use this information to develop better therapies.

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

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. The sequence of the human genome. *Science* 2001;291:1304–1351.
3. Hutchison CA III. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* 2007;35:6227–6237.
4. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;11:31–46.
5. Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. *Genome Biol* 2011;12:217.
6. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; 11:415–425.
7. Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, *et al*. Exome sequencing and the genetic basis of complex traits. *Nat Genet* 2012;44: 623–630.

8. Torgerson DG, Capurso D, Mathias RA, Graves PE, Hernandez RD, Beaty TH, Bleecker ER, Raby BA, Meyers DA, Barnes KC, *et al.* Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am J Hum Genet* 2012;90:273–281.
9. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010;7:111–118.
10. Clark MJ, Chen R, Lam HY, Karczewski KJ, Euskirchen G, Butte AJ, Snyder M. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011;29:908–914.
11. Druley TE, Vallania FL, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, *et al.* Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 2009;6:263–265.
12. Bansal V, Tewhey R, Leproust EM, Schork NJ. Efficient and cost effective population resequencing by pooling and in-solution hybridization. *PLoS ONE* 2011;6:e18353.
13. Harakalova M, Nijman IJ, Medic J, Mokry M, Renkens I, Blankenstein JD, Kloosterman W, Baas AF, Cuppen E. Genomic DNA pooling strategy for next-generation sequencing-based rare variant discovery in abdominal aortic aneurysm regions of interest—challenges and limitations. *J Cardiovasc Transl Res* 2011;4:271–280.
14. Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. *Brief Bioinform* 2010;11:484–498.
15. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;12:443–451.
16. Shendure J. Next-generation human genetics. *Genome Biol* 2011;12:408.
17. Ionita-Laza I, Makarov V, Yoon S, Raby B, Buxbaum J, Nicolae DL, Lin X. Finding disease variants in mendelian disorders by using sequence data: methods and applications. *Am J Hum Genet* 2011;89:701–712.
18. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–498.
19. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 2010;44:293–308.
20. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
21. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;34:188–193.
22. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 2010;11:773–785.
23. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;12:628–640.
24. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–1081.
25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
26. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
27. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;15:901–913.
28. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–1073.
29. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470:59–65.
30. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010;328:636–639.
31. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, *et al.* De novo mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat Genet* 2010;42:483–485.
32. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010;42:30–35.
33. Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, *et al.* Whole-genome sequencing of a single proband together with linkage analysis identifies a mendelian disease gene. *PLoS Genet* 2010;6:e1000991.
34. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for mendelian disease gene discovery. *Nat Rev Genet* 2011;12:745–755.
35. O’Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 2011;43:585–589.
36. Regalado ES, Guo DC, Villamizar C, Avidan N, Gilchrist D, McGillivray B, Clarke L, Bernier F, Santos-Cortez RL, Leal SM, *et al.* Exome sequencing identifies *SMAD3* mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms. *Circ Res* 2011;109:680–686.
37. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009;324:387–389.
38. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME, *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* 2010;42:684–687.
39. Pleasance ED, Stephens PJ, O’Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010;463:184–190.
40. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 2010;465:473–477.
41. Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, Wright FA, Rieder MJ, Tabor HK, Nickerson DA, *et al.* Exome sequencing of extreme phenotypes identifies *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nat Genet* 2012;44:886–889.
42. Todd JL, Goldstein DB, Ge D, Christie J, Palmer SM. The state of genome-wide association studies in pulmonary disease: a new perspective. *Am J Respir Crit Care Med* 2011;184:873–880.
43. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, Fingerlin TE, Zhang W, Gudmundsson G, Groshong SD, *et al.* A common *MUC5B* promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011;364:1503–1512.
44. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet* 2010;11:476–486.
45. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* 2011;7:e1002144.