AMERICAN THORACIC SOCIETY DOCUMENTS

High-Throughput Sequencing in Respiratory, Critical Care, and Sleep Medicine Research

An Official American Thoracic Society Workshop Report

Craig P. Hersh, Ian M. Adcock, Juan C. Celedón, Michael H. Cho, David C. Christiani, Blanca E. Himes, Naftali Kaminski, Rasika A. Mathias, Deborah A. Meyers, John Quackenbush, Susan Redline, Katrina A. Steiling, Holly K. Tabor, Martin D. Tobin, Mark M. Wurfel, Ivana V. Yang, and Gerard H. Koppelman; on behalf of the American Thoracic Society Section on Genetics and Genomics

This Official Workshop Report of the American Thoracic Society (ATS) was approved by the ATS Board of Directors, October 2018

Abstract

High-throughput, "next-generation" sequencing methods are now being broadly applied across all fields of biomedical research, including respiratory disease, critical care, and sleep medicine. Although there are numerous review articles and best practice guidelines related to sequencing methods and data analysis, there are fewer resources summarizing issues related to study design and interpretation, especially as applied to common, complex, nonmalignant diseases. To address these gaps, a single-day workshop was held at the American Thoracic Society meeting in May 2017, led by the American Thoracic Society Section on Genetics and Genomics. The aim of this workshop was to review the design, analysis, interpretation, and functional follow-up of high-throughput sequencing studies in respiratory, critical care, and sleep medicine research. This workshop brought together experts in multiple fields, including genetic epidemiology, biobanking, bioinformatics, and research ethics, along with physician-scientists with expertise in a range of relevant diseases. The workshop focused on application of DNA and RNA sequencing research in common chronic diseases and did not cover sequencing studies in lung cancer, monogenic diseases (e.g., cystic fibrosis), or microbiome sequencing. Participants reviewed and discussed study design, data analysis and presentation, interpretation, functional follow-up, and reporting of results. This report summarizes the main conclusions of the workshop, specifically addressing the application of these methods in respiratory, critical care, and sleep medicine research. This workshop report may serve as a resource for our research community as well as for journal editors and reviewers of sequencing-based manuscript submissions in our research field.

Keywords: bioinformatics; functional genomics; genetic epidemiology; RNA sequencing; whole-genome sequencing

ORCID IDs: 0000-0002-1342-4334 (C.P.H.); 0000-0003-2101-8843 (I.M.A.); 0000-0002-6139-5320 (J.C.C.); 0000-0002-4907-1657 (M.H.C.); 0000-0002-0301-0242 (D.C.C.); 0000-0002-2868-1333 (B.E.H.); 0000-0001-5917-4601 (N.K.); 0000-0002-2702-5879 (J.Q.); 0000-0001-9663-2093 (K.A.S.); 0000-0003-1005-5008 (H.K.T.); 0000-0002-3596-7874 (M.D.T.); 0000-0002-1018-489X (I.V.Y.); 0000-0001-8567-3252 (G.H.K.).

You may print one copy of this document at no charge. However, if you require more than one copy, you must place a reprint order. Domestic reprint orders: amy.schriver@sheridan.com; international reprint orders: louisa.mott@springer.com.

Correspondence and requests for reprints should be addressed to Craig P. Hersh, M.D., M.P.H., Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Avenue, Boston MA, 02115. E-mail: craig.hersh@channing.harvard.edu.

Ann Am Thorac Soc Vol 16, No 1, pp 1–16, Jan 2019
Copyright © 2019 by the American Thoracic Society
DOI: 10.1513/AnnalsATS.201810-716WS
Internet address: www.atsjournals.org

Contents

Overview
Workshop Agenda
Principles of Study Design
- **Study Designs and Phenotyping for Genetic Epidemiology**
- **Biobanks**
- **Health Equity**
- **Research Ethics**

Next-Generation DNA Sequencing
- **Study Design**
- **Identifying Causal Variants**

Next-Generation RNA Sequencing
- **Study Design**
- **Sequencing Methods**
- **Reporting for RNA-Seq Experiments**

Data Analysis
- **Bioinformatics**
- **Quality Control**
- **Statistical Analysis**

Data Sharing
Cell Type Heterogeneity
Single-Cell Sequencing
Multiomics Integration
Functional Validation
Conclusions

Overview

High-throughput, next-generation sequencing (NGS) technologies (*see* Box 1 for a glossary of terms) are becoming increasingly used in studies of common, complex diseases, including respiratory diseases such as asthma, chronic obstructive pulmonary disease, and idiopathic pulmonary fibrosis; critical illnesses; and sleep disorders (Table 1, Figure 1) (1–15). RNA sequencing is now more cost effective than microarrays, and exome and whole-genome DNA sequencing are rapidly replacing genotyping arrays. Given the widespread application of these techniques in respiratory, critical care, and sleep medicine research, a workshop was organized at the ATS International Conference in Washington, DC in May 2017. The aim of this workshop was to review the design, analysis, interpretation, and functional follow-up of high-throughput sequencing studies in respiratory, critical care, and sleep medicine research. Although reviews and best-practices guidelines for DNA and RNA sequencing have been published (16, 17), this workshop focused on the application of DNA and RNA sequencing to common, complex diseases in human populations but not on epigenome or microbiome studies or cancer genetics.

Workshop Agenda

The workshop participants focused on five topics, each of which concluded with a panel discussion. Areas of emphasis included study design, ethical considerations and health inequalities, applications of DNA and RNA sequencing, cell type heterogeneity, and functional studies. Biomedical literature searches were conducted by the speakers and co-chairs. The co-chairs collected summaries from speakers, and a writing group prepared the document for review by the workshop participants. Recommendations were formulated by discussion and consensus (Box 2, Figure 2).

Principles of Study Design

Study Designs and Phenotyping for Genetic Epidemiology

The general principles of epidemiology study design remain true for genetic epidemiology studies, including subject ascertainment, phenotype definition, and sample size considerations, as summarized in the Strengthening the Reporting of Genetic Association Studies (STREGA) Guidelines (18). There are several possible designs for genetics studies of respiratory disease. In the past, most studies enrolled subjects ascertained for a specific condition. These studies usually use careful phenotyping to define the disease of interest using endotypes, such as methacholine challenge testing or polysomnography (19–21). General population (cohort) studies may offer the advantage of large sample sizes and the ability to study multiple outcomes, although the phenotyping may not be as precise. Questionnaires may be the primary source of respiratory disease diagnosis,

Box 1. Definitions of Commonly Used Terms in Sequencing Studies

Batch effect: In a large study, library construction and sequencing is done in batches (e.g., 96-well plate), which is a source of technical variation that should be addressed in the data analysis.

Complex trait: A disease or phenotype that does not follow Mendelian inheritance. Complex traits are likely influenced by multiple genes and environmental factors. Most common human diseases would be considered complex diseases.

Deconvolution: RNA sequencing is frequently performed in tissues such as blood or lung, which are composed of multiple cell types. Deconvolution methods aim to estimate the cell type proportions and/or identify the cell type(s) responsible for the expression of specific genes.

Exome: The portion of the human genome (approximately 1%) that encodes for proteins. Whole-exome sequencing (WES) specifically targets these sequences.

Expression quantitative trait locus (eQTL): A genetic variant, usually an SNP (*see below*), that affects the expression of a gene. eQTLs can be located near the gene of interest (cis-eQTL) or distant (>1 Mb away) (trans-eQTL).

Genome-wide association study (GWAS): A study that assays hundreds of thousands to millions of SNPs across the genome and tests each variant for association with a disease or trait of interest.

Mendelian disorder: A disease determined by variation in a single gene (e.g., cystic fibrosis or sickle cell disease).

Next-generation sequencing (NGS): Highly automated parallel sequencing technique of small fragments of DNA or RNA. Millions or even billions of nucleotides, up to a whole genome, can be determined in 1 day.

RNA integrity number (RIN): A proprietary algorithm that quantifies RNA degradation on the basis of an electropherogram.

Single-nucleotide polymorphism (SNP): A single base pair change in DNA sequence (e.g., C to T) which is prevalent (>1%) in the general populations. SNPs are the most common type of genetic variation.

Sequencing coverage/depth: For DNA sequencing, the number of reads that include a specific nucleotide in the sequencing experiment. This can be averaged across the genome (e.g., 30×). For RNA sequencing, sequencing depth is usually presented as the total number of sequencing reads.

Variant calling: The process of identifying genetic variants (usually SNPs) in an individual exome or genome sequence.

Table 1. Examples of human next-generation sequencing studies in respiratory, critical care, and sleep medicine

Technology	Disease/Trait	Study Design/ Subjects	Main Findings	Validation	Reference
Whole-exome sequencing	Narcolepsy	18 Families	8 Missense variants in *P2Y11*	Resequencing in 250 cases, 150 control subjects; *in vitro* P2Y11 signaling assays	10
Whole-exome sequencing	Bronchopulmonary dysplasia	50 Twin pairs, including 51 BPD cases	258 Genes with rare nonsynonymous mutations	Lung gene expression in published human data and rat BPD model, mouse phenotype database	11
Whole-exome sequencing	Airflow obstruction	100 Heavy smokers with normal lung function	Nonsynonymous SNP in *CCDC38*	Association testing in two additional studies. Immunohistochemistry in bronchial epithelial cells.	6
Whole-exome sequencing	Idiopathic pulmonary fibrosis	79 Probands with familial pulmonary fibrosis, 2,816 control subjects	Mutations in *PARN* found in cases, not control subjects. Mutations in *RTEL1* more common in cases vs. control subjects	Mutations segregated in families. Shorter leukocyte telomeres in mutation carriers.	8
Whole-genome sequencing	Pulmonary vascular disease	864 PAH, 16 PVOD/ PCH, 7,134 control subjects	*EIF2AK4* mutations in 19 patients with PAH	Phenotype association with younger age, reduced KCO, shorter survival	12
Whole-genome sequencing	Asthma	WGS in 8,453 Icelanders, imputation in >150 K	Rare variant in *IL33* associated with lower eosinophil count, reduced asthma risk	Genotyping in 6,465 cases, >300 K control subjects; interleukin-33 gene expression; *in vitro* assay of receptor binding	3
RNA sequencing	Smoking	Blood samples from 229 current, 286 former smokers	171 DE genes, including 7 lncRNAs, 8 genes with differential exon use	Published microarray study	13
RNA sequencing	COPD	Lung tissue from 98 cases, 91 control subjects	2,312 DE genes	qPCR for seven genes	2
Single-cell RNA sequencing	IPF	FACS-sorted lung epithelial cells from 6 IPF, 3 control subjects	4 Cell clusters: AT2, basal, goblet, and indeterminate	Immunofluorescence confocal microscopy for epithelial cell markers	15
miRNA sequencing	Sepsis	Plasma from 29 sepsis, 44 noninfective SIRS, 16 control subjects	6 miRNAs distinguish sepsis from SIRS	qPCR, correlation with inflammatory cytokines	9
miRNA sequencing	Exercise physiology	Plasma before/ after treadmill exercise test, $n = 26$	miR-181b increased with exercise	qPCR in separate cohort ($n = 59$), Skeletal muscle expression in mouse exercise model	14

Definition of abbreviations: BPD = bronchopulmonary dysplasia; COPD = chronic obstructive pulmonary disease; DE = differentially expressed; FACS = fluorescence-activated cell sorter; IPF = idiopathic pulmonary fibrosis; KCO = carbon monoxide transfer coefficient; lncRNA = long noncoding RNA; PAH = pulmonary arterial hypertension; PCH = pulmonary capillary hemangiomatosis; PMVEC = pulmonary microvascular endothelial cells, PVOD = pulmonary veno-occlusive disease; qPCR = quantitative polymerase chain reaction; SIRS = Systemic Inflammatory Response Syndrome; SNP = single-nucleotide polymorphism; WGS = whole-genome sequencing.

although several large cohorts have included spirometry (22, 23). For common diseases, the large numbers may offset the potential for phenotypic heterogeneity (e.g., childhood vs. adult-onset asthma) or even misclassification (e.g., chronic obstructive pulmonary disease [COPD] misdiagnosed as asthma, especially in women) (24). Recent studies have linked genetics to the electronic medical record (25). General population cohorts have limited utility in studies of critical illness, where subjects are enrolled in the hospital (26). Although most recent studies have been case–control or cohort studies, family-based studies still play a role, especially in the analysis of rare variants, where transmission can be followed through a pedigree (27).

As genomic data sharing has become the norm, secondary analysis for respiratory diseases is now routinely performed in general population studies, such as the Framingham Heart Study (22). When secondary data will be used or when multiple studies will be combined in a meta-analysis, investigators must carefully review the phenotyping methods, including

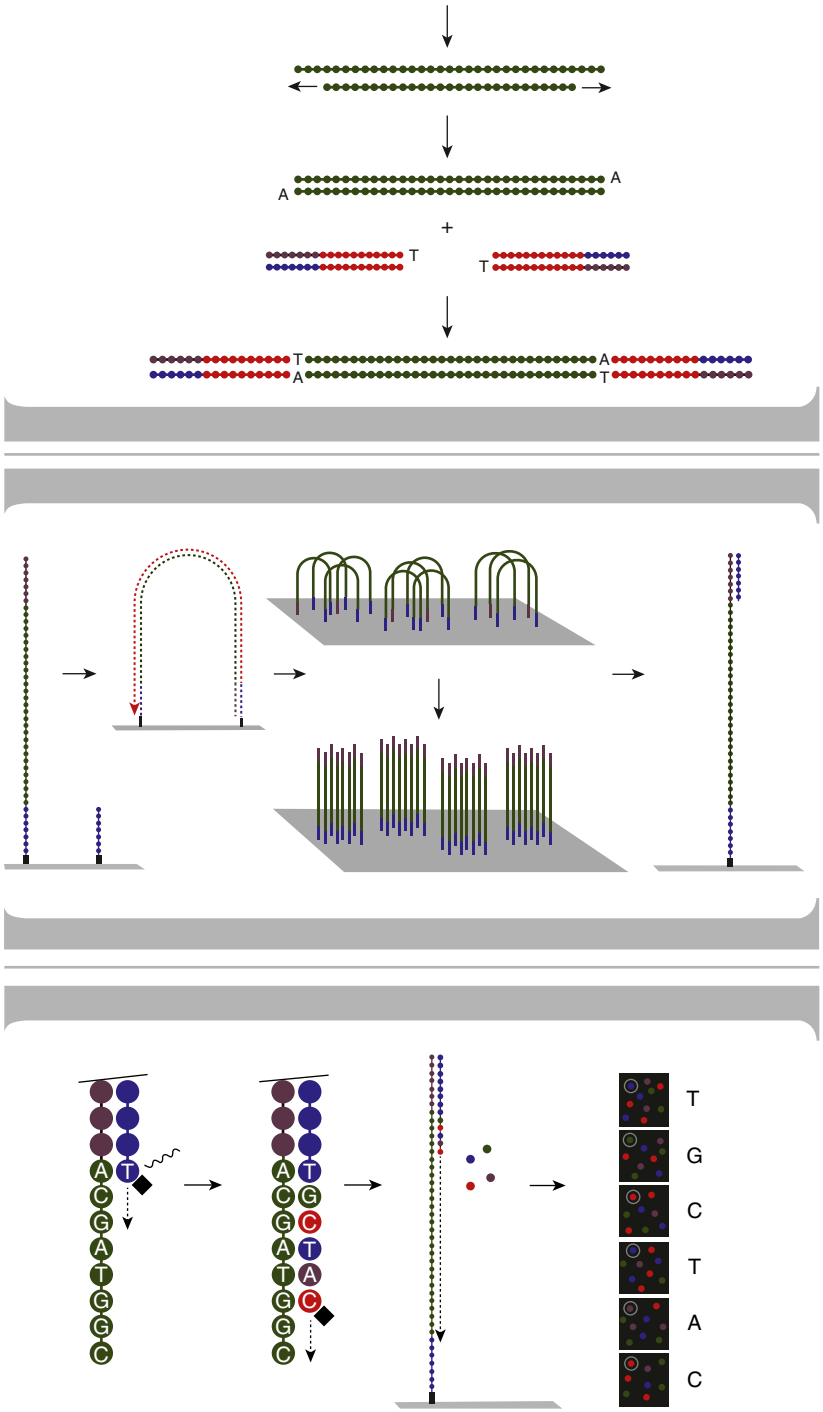

Figure 1. Next-generation sequencing methodology (Illumina). Genomic DNA is fragmented and sequencing adaptors are attached. The genomic library is then hybridized to complementary oligonucleotide probes in the flow cell chamber. Because there are adaptors on both ends, hybridization results in a bridge. Amplification leads to clusters of fragments with the same sequence. Clusters are denatured; then, sequencing-by-synthesis involves the addition of fluorescently labeled nucleotides, with serial imaging after the incorporation of each nucleotide. Reprinted by permission from Reference 116.

the specific questionnaire items, to be sure that similar traits are being compared. In case–control studies, control subjects should have comparable exposures, such as smokers with normal lung function in COPD studies or patients with multitrauma, pneumonia, or sepsis who did not develop acute respiratory distress syndrome (28–30).

Biobanks

Sequencing studies undertaken at the scale required for well-powered association testing require organized efforts, with coordinated biobanking. Rare diseases and phenotypes may be due to genetic variants with high penetrance, which may be detectable in relatively modest numbers of samples. Even in this situation, and in the absence of many different mutations causing similar phenotypes, it is helpful to draw on very large numbers of sequenced individuals. Genomics England (the "100,000 Genomes Project") (31), the U.S. National Heart, Lung, and Blood Institute Trans-Omics in Precision Medicine (TOPMed) (32), and the Genome Aggregation Database (33) are initiatives that will enhance such comparisons, which are critical to confirm whether variants are causal for the disease in question (Table 2). The Genomics England project is recruiting patients with cancers, infectious diseases such as tuberculosis, and rare diseases, including primary ciliary dyskinesia, spontaneous pneumothorax, familial pulmonary fibrosis, and familial multiple pulmonary arteriovenous malformations.

Complementing efforts that specifically recruit individuals with particular diseases are population biobanks that are agnostic to health status, many of which have extensive longitudinal follow-up. The UK Biobank recruited 500,000 participants aged 40 to 69 years (34). In addition to a baseline assessment, subsequent health status is evaluated via linked electronic healthcare records. Beginning with respiratory studies in 50,000 participants (35), genome-wide genotyping has now been extended to all participants. Whole-exome sequencing (WES) is underway, and whole-genome sequencing (WGS) has recently been announced in collaboration with industry partners. The sequence data will be made available to the research community. Although 95% of the UK Biobank is of European ancestry, similar efforts are in progress in China in the Kadoorie Biobank (36).

Health Equity

Many respiratory, critical care, and sleep disorders have substantial differences in disease susceptibility, prevalence, and burden according to race and ethnicity (37). Genetic and genomic factors, along with their interplay with the environment, contribute to these differences. For example, African ancestry is a strong predictor of lung

Box 2: Recommendations for Design and Analysis of Next-Generation Sequencing Studies

- General principles of genetic epidemiology study design are especially important in next-generation sequencing studies. Disease-specific studies may have more detailed phenotyping, whereas population studies may allow for larger sample sizes.
- Investigators must clearly define the phenotypes of both cases and control subjects, including consideration of relevant exposures, such as smoking.
- Relying only on datasets largely composed of individuals of European ancestry will limit discoveries, especially in diseases that may be more prevalent in other racial or ethnic groups. Therefore, we suggest expanding sequencing efforts in subjects of different ancestries.
- Studies should consider broad consent for secondary data use.
- Researchers should use standardized methods of experimental design and data analysis for exome and whole-genome sequencing studies.
- Methods for design and analysis in RNA sequencing studies are more variable. Researchers should clearly document their methods, including software versions and input parameters, and consider validating key results with a different analysis method.
- Multidisciplinary teams should include bioinformaticians, statisticians, and computational biologists to assist in the management and analysis of large datasets.
- Quality control is the responsibility of the investigators. It should not be assumed that the core sequencing facility has performed all the necessary quality control steps.
- Data sharing is required by many funders and should be the default.
- Because of the extreme cellular heterogeneity of the lung, future studies should address cellular heterogeneity in the design and analysis by any of the proposed methods.
- We anticipate an important role of single-cell sequencing in the future. For wider acceptance, single-cell sequencing has to address specific hypotheses and should be held to the same rigorous standards as other study designs, even while the laboratory and statistical methods are under development.
- Given the pervasive influence of circadian biology on multiple cellular processes, including gene expression, studies should time stamp sample collection.
- Integrating different omics datasets, both at the single-cell and at the tissue level, will likely increase our understanding of the complex diseases in respiratory, critical care, and sleep medicine.
- Laboratory validation is an important next step toward the eventual translation of results.

function (38). Some disease susceptibility or pharmacogenetic variants that have been identified in diseases such as asthma, emphysema, and sleep apnea (39–41) are population specific and may be absent or low frequency in other populations (42, 43). Although most of these studies have been performed using common variants, there is even more ethnic diversity for rare variants (44). Individuals of African ancestry harbor a larger number of rare variants than white individuals, which may have important clinical implications. For example, rare variants identified as causing cardiomyopathy in white individuals were so common in African Americans as to indicate that they were unlikely to be pathogenic (45). To date, there has been a large gap in research studies involving non-white individuals, for reasons including convenience, access, and genetic heterogeneity (46). In addition, there have also been disparities in funding minority investigators or diseases that predominately affect minorities, such as sickle cell disease (47). By 2060, only 44% of the United States will be non-Hispanic white (48). There is a strong scientific and moral justification for expanding sequencing studies into other ancestries. Efforts such as TOPMed are performing WGS in a large number of non-European samples. These and other efforts can help ensure that sequencing efforts improve health equity for the benefit of all.

Research Ethics

Several ethical challenges have emerged related to NGS studies. Because of the ability to assay numerous sites and the requirements for data sharing by the U.S. National Institutes of Health (NIH) and other funders (49), genome-wide association studies and sequencing data are often used for secondary studies, which may be unrelated to the initial trait or disease proposed. To allow for these studies, investigators must request and subjects must provide broad consent for secondary data analysis, as opposed to narrow consent for a specific disease. There is no consensus about what is required for broad consent for databases such as the National Center for Biotechnology Information database of Genotypes and Phenotypes (dbGaP) (50); these decisions are frequently left to local institutional review boards.

In addition to the genes of interest, WES and WGS studies will identify other genetic variants that may be clinically significant for the subject or their family members. The American College of Medical Genetics has provided recommendations for the reporting of secondary results from clinical sequencing (51), providing a list of 59 actionable genes, which, interestingly, does not include the genes for cystic fibrosis or alpha-1 antitrypsin deficiency. However, it is not clear how these recommendations would apply to research studies, where the sequencing is not performed in a Clinical Laboratory Improvement Amendments (CLIA)-certified laboratory. There is usually no mechanism or resources for confirmatory clinical sequencing or genetic counseling. Sequencing studies of DNA of patients with a critical illness require consent from patient proxies as well as the designation of an individual to receive study results if the patient remains incapacitated or dies (52).

Next-Generation DNA Sequencing

Study Design

Humans carry an extraordinary amount of genetic diversity. Although most variants in a given individual are common, most genetic variants in a population are rare (i.e., present in <1% of the population).

Personalis EX2019

Table 2. Biobanks and commonly used databases for next-generation sequencing research

	URL
Biobanks and other large sequencing studies	
Centers for Common Disease Genomics	www.genome.gov/27563570
China Kadoorie Biobank	www.ckbiobank.org
Genomics England ("100,000 Genomes Project")	www.genomicsengland.co.uk
Trans-Omics in Precision Medicine (TOPMed)	www.nhlbiwgs.org
U.K. Biobank	www.ukbiobank.ac.uk
Databases	
Database of Genotypes and Phenotypes (dbGaP)	www.ncbi.nlm.nih.gov/gap
Ensembl genome browser	www.ensembl.org
Gene Expression Omnibus (GEO)	www.ncbi.nlm.nih.gov/geo
Genome Aggregation Database (gnomAD)	http://gnomad.broadinstitute.org/
Genotype-Tissue Expression project (GTEx)	www.gtexportal.org
Human Cell Atlas	www.humancellatlas.org
Lung Map	www.lungmap.net
Reference Sequence Database (RefSeq)	www.ncbi.nlm.nih.gov/refseq
Sequence Read Archive (SRA)	www.ncbi.nlm.nih.gov/sra
University of California Santa Cruz (UCSC) Genome Browser	www.genome.ucsc.edu

Genotyping microarrays, together with methods of genotype imputation, efficiently allow testing of more prevalent genetic variants. However, comprehensive assessment of DNA variation, including discovery of rare or novel variants, requires a test that assays each base pair. The exponential decrease in sequencing costs has led to the ability to perform WGS studies to identify the contributions of rare variants to disease.

The design of a DNA sequencing study should consider several factors. Sequencing depth determines the accuracy of variant calling in an individual. For example, 30× coverage leads to high accuracy over most of the genome. However, one can sequence more samples for the same cost using lower coverage (e.g., 3–6×). Although less accurate, lower coverage still provides suitable variant information for genetic association testing and may be superior to genotyping (53). Another consideration is targeted (e.g., exome) versus WGS. The exome harbors most rare, highly deleterious mutations; sequencing only these regions reduces costs, although these savings are offset somewhat by the additional costs and inefficiencies of library preparation. Some coding regions may still be poorly covered by WES (54).

Identifying Causal Variants

NGS has led to breakthroughs in the discovery of genes for Mendelian diseases and other rare variants of strong effect (55, 56). Several groups have provided guidelines for identifying causal variants for Mendelian disease (57) and for identifying genetic association in WES (58) and WGS (59). However, the interpretation of WES and WGS data has specific challenges. Very small error rates over billions of base pairs have the potential to generate many false positives (60), although advances in technology, approaches, and bioinformatics methods have vastly improved data quality. Similarly, the large number of variants carried by any individual (27) can also lead to false positives, and caution must be used to avoid inflated estimates of pathogenicity (61). In addition, most studies of rare variants are likely underpowered (5, 58, 62). The growing availability of population-specific high-quality reference genomes will aid comparison of diseased study populations to these reference datasets. Coordinated efforts are providing population-based reference data important for filtering causal variants (33) and performing WGS in large numbers of subjects, such as the NIH Centers for Common Disease Genomics and TOPMed. These efforts will lead to new rare variant discoveries as well as improved reference panels for genotyping studies and fine mapping. Table 3 details recommendations for reporting the results of a WES or WGS study.

Next-Generation RNA Sequencing

Study Design

In contrast to genome sequence, gene expression is dependent on cell, tissue, and disease state. Although this context dependence makes sample collection and assays more challenging, gene expression may be more closely reflective of disease pathophysiology. Gene expression microarrays have been largely replaced by RNA-seq, which can assay a broader range of RNA types with increased sensitivity and lower costs. RNA-seq studies can be grouped into two broad categories on the basis of their study designs (63). Annotation studies aim to define the transcriptome of a specific cell type or organism, including novel transcripts. In comparison, quantification or differential expression studies compare transcript levels across experimental conditions or diseases. An introduction to RNA-seq for bench science has been recently published (64).

Several factors are important in designing an RNA-seq study in human populations. Because sequencing costs depend on the number of reads, there is an inherent trade-off between sample size and sequencing depth, similar to DNA sequencing. As low as 1 to 10 million reads per sample may be adequate for differential expression (65, 66), whereas up to 200 million reads may be required to define all isoforms (67, 68). For most human disease studies, the number of samples may be more important than the number of reads (69).

Sequencing Methods

Sequencing technology has been reviewed elsewhere (70, 71). Library construction follows one of three general protocols (72). Poly-A capture is best for selecting coding transcripts but requires the highest-quality input RNA. Total RNA libraries include more RNA species; ribosomal RNA depletion is used to improve yield. Methods to capture specific target sequences can be used for lower-quality or fragmented RNA, including formalin-fixed paraffin-embedded tissue. Globin depletion is a common step for whole blood samples. Small RNA sequencing (i.e., microRNA) requires a specific library construction protocol.

Investigators must determine the minimum quality for input RNA, on the basis of RNA integrity number. Other

Personalis EX2019

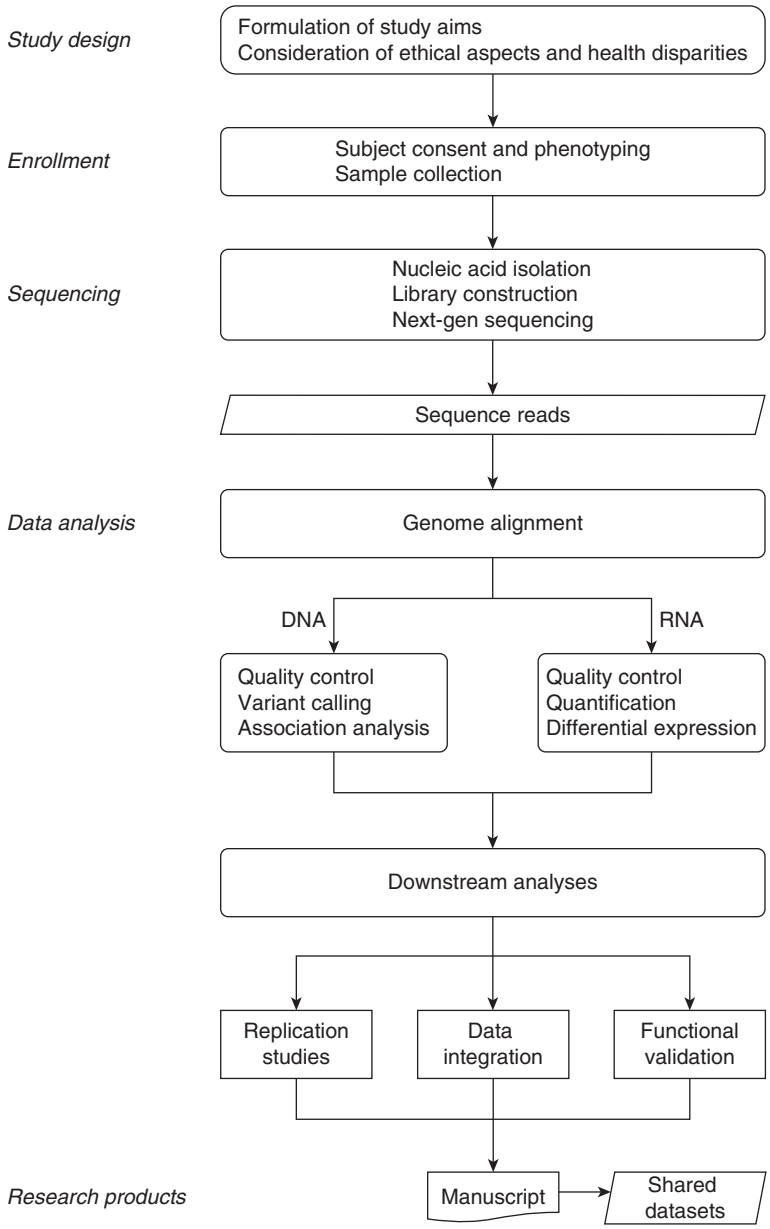

Figure 2. Workflow for a next-generation sequencing study in human disease.

variables in sequencing include stranded versus unstranded protocols, single versus paired end reads, and insert length. Single-end short reads (≤75 bp) are acceptable for standard differential expression (73). Paired-end longer reads are preferable for annotating splicing events. In larger studies, samples should be randomized across batches to minimize confounding; analysis should account for batch effects. Gene expression in human tissues can be quite variable, depending on sampling, technical factors, and cellular heterogeneity (74). The latter can be addressed by deconvolution methods, which require additional information, either cell counts or cell-specific reference transcriptomes.

A review of the specific steps and analytic tools for RNA-seq data analysis has been recently published (16). However, there remains controversy regarding the optimal methods for normalization of RNA-seq data and for detecting differential expression. A few studies have examined these questions systematically (75–79). Furthermore, selection of an analytic approach is also dependent on the experimental design; for example, not all available tools support inclusion of covariates. Caution is advised regarding the interpretation of results where the number of biologic replicates in an experiment is small or where transcripts are expressed at very low levels.

Reporting for RNA-Seq Experiments

We propose minimum elements to be reported for RNA-seq experiments (Table 3), which extend the Minimal Information about Microarray Experiments (MIAME) guidelines (80). These minimal reporting requirements are important because of the rapidly evolving landscape of methods and tools for the analysis of RNA-seq data, the need to ensure RNA-seq data are both interpretable and reproducible, and the need to facilitate access to and integration of RNA-seq experiments across a spectrum of biologic and experimental conditions. Given the lack of consensus on the optimal methods for mapping versus assembling RNA-seq reads, normalizing RNA-seq data, and assessing for differential expression, we encourage investigators to repeat key analyses using more than one approach.

Data Analysis

Bioinformatics

Sequencing technologies have improved to the point where the greatest barrier to obtaining scientific insights is more related to data storage, analysis, and interpretation than its generation (81). The first critical component is an interdisciplinary team with expertise encompassing both the design and the use of specialized methods on sophisticated computational resources (82, 83). Institutional infrastructure or external service providers that offer high-performance computing environments, including cloud computing and core facilities, are important to facilitate the generation and analysis of high-throughput sequence data. One significant advantage to these solutions is that they distribute costs over many users. In addition, these resources can help ensure high-quality data and results, as they are generated by devoted personnel who are more familiar with NGS approaches than occasional users. On the other hand, some analytic steps are best performed with feedback from those familiar with experimental design, rather than by pipelines that may overlook

Personalis EX2019

Table 3. Minimal elements required in the reporting of high-throughput sequencing studies.

Analytic Step	Required Elements
Whole-exome and genome sequencing	
Preprocessing and preanalysis quality control	Randomization of samples Target design, when applicable (e.g., whole-exome sequencing) Methods for quality assessment of: Raw reads Aligned reads and coverage Global data quality Ancestry of samples (comparison with study and to reference genomes)
Core analytics	Method of read alignment Method of variant calling Method of association analyses
Advanced analytics	Methods for integration with other data types
RNA sequencing	
Preprocessing and preanalysis quality control	Spike-in use Randomization of samples Number of raw reads Methods for quality assessment of: Raw reads Aligned reads Quantification of reads Reproducibility of replicates Global data quality
Core analytics	Method of transcript/gene identification Method of transcript/gene quantification Method of normalization Method of batch correction Method of detection of differential expression
Advanced analytics	Method of transcript/isoform discovery Method of indel detection Method of gene fusion detection Method of variant detection Method for single-cell analyses Methods for integration with other data types

Required elements should also include the package or software name, version number, and settings used for the analysis.

important issues. Ideally, researchers with *1*) backgrounds in chemistry and molecular biology involved in generating the data, *2*) in-depth familiarity with analysis of sequencing data, and *3*) backgrounds in design of a particular experiment will communicate intermediate results and tailor analytical steps as needed.

Quality Control

Quality control (QC) in a rigorous and standardized matter is critical. After raw reads are demultiplexed after their generation by the sequencing instrument, QC steps to ensure sequencing worked appropriately provide output, including number of reads per sample, quality score of reads at each base, and overrepresented sequences (e.g., primers). Subject-level QC includes checking for sex consistencies, duplicates, and related subjects. After reads are aligned to a reference genome, additional parameters assess mapping quality. For example, software tools (Tables 4 and 5) can be used to summarize the number of mapped reads, including junction-spanning reads for RNA-seq, and can compute the number of bases assigned to various classes of DNA/RNA according to a reference file. For RNA-seq, use of Ambion External RNA Controls Consortium spike-ins offers an additional QC measure. Scripts that provide standardized and reproducible reports on the basis of output from these various programs, such as those that are part of the Genome Analysis Toolkit (GATK) best-practices pipeline for DNA-Seq (84) or taffeta scripts for RNA-seq (85), facilitate the assessment of QC and decisions about which samples should be excluded or what kinds of bias might be present. Mapped read files that are in bam format can be converted to bigwig or other compressed format for display in a browser, such as the University of California Santa Cruz (UCSC) genome browser, to verify that mapping of particular genes looks as expected (e.g., that full lengths of genes are covered vs. highly irregular portions).

Statistical Analysis

After alignment of reads, DNA-Seq data are analyzed to search for variation relative to a reference genome, and Variant Call Format (vcf/bcf) files are obtained. There are a variety of statistical tests for association within the WGS framework. The simplest consideration is based on frequency of the variant and/or a prior probability of being associated with disease (e.g., known pathogenic, deleterious, functionally relevant). Common variants, defined as those with minor allele frequency greater than 0.01 to 0.05 or minor allele count greater than 10 to 20, depending on the sample size and factors such as case–control imbalance, are usually analyzed with a genome-wide association study approach of single-nucleotide polymorphism (SNP) association tests (86). Rare and/or deleterious variants are generally analyzed using a window-based approach, where windows consist of SNPs in or near a gene or are based on sliding genome regions of 5 to 50 kb. Statistics for each window are obtained using different approaches: burden tests collapse many variants into a single risk score but assume all variants are similar in effect (i.e., risk alleles); adaptive burden tests collapse many variants but allow for risk, protective and neutral; and variance component tests apply random effects modeling (87). There is also a class of window-based rare variant tests that combine the variance component and burden test framework to take advantage of strengths of both; sequence kernel association test-optimal (SKAT-O) is the most commonly used (88). It is routine to apply multiple rare-variant tests and use alternative options of windows, resulting in increased multiple testing complexity. In addition to Bonferroni correction,

Table 4. Software for DNA sequencing studies

Task	Tools	URL
Alignment	BWA-MEM (117)	http://bio-bwa.sourceforge.net
	Bowtie2 (118)	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Quality control	Raw reads	
	FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
	Mapping	
	BAMtools (119)	https://github.com/pezmaster31/bamtools/
	Picard Tools	https://broadinstitute.github.io/picard/
	Variants	
	GATK (120)	https://software.broadinstitute.org/gatk/
Variant calling	SAMtools (121)	http://www.htslib.org
	GATK unified genotyper, haplotype caller, variant quality score recalibration (122)	https://software.broadinstitute.org/gatk/
Visualization	Integrative Genomics Viewer (IGV) (123)	http://software.broadinstitute.org/software/igv/
	UCSC Genome Browser (124)	http://www.genome.ucsc.edu/
Association analysis	PLINK 2 (common variants) (125)	https://www.cog-genomics.org/plink2/
	SKAT-O (rare variants) (88)	https://www.hsph.harvard.edu/skat/
	GENESIS (rare variants) (126)	https://bioconductor.org/packages/release/bioc/html/GENESIS.html
	BOLT-LMM (127)	https://data.broadinstitute.org/alkesgroup/BOLT-LMM/

This table provides an overview of commonly used software tools for performing analysis of next-generation sequencing data. Because the field continues to evolve rapidly, additional tools not listed in this table may also be useful to researchers.

permutation and Bayesian approaches are used to adjust for multiple comparisons.

Controversy about the best way to analyze RNA-seq data still exists, and methods development is ongoing (89). For samples passing initial QC, the next step involves quantification of levels of genes or transcripts. In most cases, reference gene or transcript files are obtained from Ensembl (90) or RefSeq (91). The output of the quantification process is then used with an appropriate software package to measure differential expression and assess related QC. Regardless of program used, it is important to report false-discovery rate, adjusted *P* values and fold changes.

Data Sharing

To ensure that high-throughput sequence results are reproducible and that costly data can benefit all stakeholders, data-sharing resources have grown significantly. Both raw data and results generated from projects sponsored by major funders are required to be deposited into publicly available databases. DNA-Seq data, including that of TOPMed, are deposited in dbGaP, along with individual-level phenotype and association results (50). RNA-seq and other sequencing data are available in the Sequence Read Archive and can be discovered via the Gene Expression Omnibus (92).

Cell Type Heterogeneity

An important issue for consideration in many omics studies of the lung is cell heterogeneity. The lung is a complex organ comprising approximately forty resident cell types (93), a growing number of cell subpopulations that are present either transiently during development or in adult lung (94), as well as many types of inflammatory cells that infiltrate the airways and alveoli during periods of injury or disease. Thus, a signal measured by omics technologies in the whole lung can reflect a change in the pattern of expression of the molecules measured within a certain cell type, a change in the cellular composition of the lung, or both. There are three main approaches to deal with cellular heterogeneity. One approach is to perform statistical deconvolution of omic profiles by relying on cell-specific features from reference datasets. This approach has been used widely in peripheral blood profiling studies (95) and more recently on complex tissues (96), but it is highly dependent on known markers and difficult to implement for lung cell populations because of the limitations of appropriate reference datasets. The second approach is to isolate cell types on the basis of cell surface markers using flow cytometry or specific areas of the lung by laser capture microdissection (LCM). Although cell sorting is often used in immunological studies and has facilitated major contributions to the field (97), it is limited by the need for known cell markers and antibodies for cell populations of interest, as well as concerns that stress from cell sorting may affect gene expression patterns. LCM can be technically challenging on human lung tissue but has had some success in conjunction with sequencing technologies (98). However, in most benign tissues, the resolution of LCM allows for enriching for a regional microenvironment but not for dissecting between different cell types.

Single-Cell Sequencing

The recently developed single-cell technologies provide the best solution to identification of all relevant cell populations, although technical limitations remain (99–101). The reproducibility and success of such studies depend greatly on availability of high-quality human tissue, on cellular susceptibility to stress, and on the platforms used (102). Despite these limitations, single-cell RNA-seq studies have shed light on lung cell population heterogeneity during lung development (103) and in lung diseases such as idiopathic pulmonary fibrosis (15). The recent development of single-nucleus RNA-seq may address some of the issues

Table 5. Software for RNA sequencing studies

Task	Tools	URL
Alignment	Bowtie (128)	http://bowtie-bio.sourceforge.net/index.shtml
	STAR (129)	https://github.com/alexdobin/STAR/
	TopHat (130)	https://ccb.jhu.edu/software/tophat/index.shtml
Transcript quantification	Cufflinks (130)	http://cole-trapnell-lab.github.io/cufflinks/
	eXpress (131)	https://pachterlab.github.io/eXpress/
	HTSeq-count (132)	http://htseq.readthedocs.io/en/master/count.html
	Kallisto (133)	https://pachterlab.github.io/kallisto/
	RSEM (134)	https://github.com/deweylab/RSEM
Quality control	Raw reads	
	FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
	Mapping	
	BAMtools (119)	https://github.com/pezmaster31/bamtools/
	Picard Tools	https://broadinstitute.github.io/picard/
	RSeQC (135)	http://rseqc.sourceforge.net
	Quantification	
	NOISeq (136)	https://bioconductor.org/packages/release/bioc/html/NOISeq.html
Differential expression	DEGseq (137)	https://bioconductor.org/packages/release/bioc/html/DEGseq.html
	DESeq2 (138)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
	edgeR (139)	http://bioconductor.org/packages/release/bioc/html/edgeR.html
	limma/voom (140)	https://bioconductor.org/packages/release/bioc/html/limma.html
	PoissonSeq (141)	https://cran.r-project.org/web/packages/PoissonSeq/index.html
	NOISeq (136)	https://bioconductor.org/packages/release/bioc/html/NOISeq.html
	Sleuth (142)	https://pachterlab.github.io/sleuth/
Alternative splicing	CuffDiff2 (143)	http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/
	DEX-Seq (144)	http://bioconductor.org/packages/release/bioc/html/DEXSeq.html
	DSG-Seq (145)	http://bioinfo.au.tsinghua.edu.cn/software/DSGseq/
	MISO (146)	http://genes.mit.edu/burgelab/miso/
	rSeqDiff (147)	http://www-personal.umich.edu/~jianghui/rseqdiff/
	Leafcutter (148)	https://github.com/davidaknowles/leafcutter
Visualization	CummeRbund (130)	http://compbio.mit.edu/cummeRbund/
	Integrative Genomics Viewer (IGV) (123)	http://software.broadinstitute.org/software/igv/
	RNASeqViewer (149)	http://bioinfo.au.tsinghua.edu.cn/software/RNAseqViewer/
	SplicePlot (150)	http://montgomerylab.stanford.edu/spliceplot/index.html
	SpliceSeq (151)	https://bioinformatics.mdanderson.org/main/SpliceSeq:Overview
	SplicingViewer (152)	http://bioinformatics.zj.cn/splicingviewer/
	UCSC Genome Browser (124)	http://www.genome.ucsc.edu

This table provides an overview of commonly used software tools for performing analysis of RNA sequencing data. Because the field continues to evolve rapidly, additional tools not listed in this table may also be useful to researchers.

with tissue quality (104). To address cellular heterogeneity in human lung disease, the field needs a "lung (disease) atlas," such as the one proposed by the human cell atlas (105), a large collaborative set of studies that will systematically profile all cells in diseased and healthy human lungs.

Multiomics Integration

DNA-seq offers the opportunity to detect different sources of DNA variation, including common and rare single-nucleotide variant and small and large deletions and insertions, whereas RNA-seq provides full assessment of the cell or tissue transcriptome at a given point in time, with a high dynamic range of these transcripts, different mRNA transcript isoforms, as well as different classes of mRNA and noncoding RNAs (ncRNAs). Integrative genomics methods evaluate the functional significance of DNA variation on gene expression. First, these address the relation of genetic variation with transcript abundance (expression quantitative trait locus, eQTL). Because SNP variants can be called in RNA-seq, this offers a direct assessment of an eQTL effect if an SNP is present in its heterozygous form; comparison of the allelic expression demonstrates the relation of an SNP with transcript abundance (allelic imbalance). Because many disease SNPs exert their function through eQTLs, well-powered lung-relevant eQTL datasets are needed. Until now, many studies have been performed in mixed cell populations of white blood cells (106) or whole lung (107). Because eQTLs may be cell specific, lung cell eQTL maps are needed to increase our understanding of lung-specific disease SNPs. Second, RNA-seq enables assessment of the relation of genetic variation with splicing events leading to alternative isoform expression (splice QTL). Finally, the association of genetic variation with transcripts induced by disease or specific stimuli (context-dependent eQTLs) can be investigated either in paired observational human studies or *ex vivo* in laboratory settings (Table 6). Because most respiratory diseases develop as a result from environmental exposures and genetic background, the study of inducible eQTLs may offer a good model to understand disease development by integrating genetic variation with induced gene expression.

Using study-specific subsets of RNA-seq data or external reference datasets such as the GTEx (Genotype-Tissue Expression project) consortium (108) allows for the ability to impute gene expression in large numbers of individuals for whom only genetic variant data are available (109). These integrative approaches can expand

Table 6. Selected examples of omics integration and using omics for functional validation studies

Technique	Example
Context-dependent eQTLs	Li and colleagues showed that cytokine production by peripheral blood mononuclear cells on stimulation depends on six specific SNPs (153). One inducible cytokine QTL at the NAA35-GOLM1 locus markedly modulated interleukin-6 production in response to multiple pathogens and also showed association with susceptibility to candidemia.
Imputed gene expression (PrediXcan)	Ferreira and colleagues tested for associations between asthma and 17,190 genes found to have cis- and/or trans-eQTLs across 12 cell types relevant to asthma (154). They confirmed 37 genes where the association was driven by eQTLs located in established risk loci for allergic disease and discovered 11 novel genetic associations.
Gene knockdown	Dixit and colleagues investigated the effect of gene knockdown by CRISPR/Cas9 on RNA-seq expression in human LPS stimulated bone marrow dendritic cells, a method they called Perturb-seq (155). By analyzing the transcriptional consequences of perturbations of transcription factors in these cells, they were able to interpret the functional consequences of these transcription factors, as well as their interaction, uncovering their molecular mechanisms.

Definition of abbreviations: eQTL = expression quantitative trait locus; LPS = lipopolysaccharide; QTL = quantitative trait locus; SNP = single-nucleotide polymorphism.

the value of smaller sample sizes with transcript data to the larger datasets to identify gene expression correlated with phenotype (Table 6).

We anticipate that further integration of other omics data in bulk tissue or at the single-cell level, through efforts such as the Lung Map (110), will markedly increase our understanding of respiratory disease. The efficiency and speed of these types of analysis may be improved by the implementation of composite measures (111) in future research programs in respiratory medicine, particularly as the approach can be used for all types of omics analysis, including transcriptomics, proteomics, metabolomics, and epigenetics. Omics integration analysis has advanced considerably, and many machine learning methods are now being used, including Bayesian and network-based approaches (112), and, more recently, deep learning and neural networks (113).

Functional Validation

In functional studies, gene expression may also be used as outcome, either *in vivo* when human subjects or patients are exposed to an environmental stimulus or drugs or in human samples or animal models. Comparative analysis of humans and mouse models through RNA-seq may enable swift validation of downstream targets and provide insight in the validity of the animal model (114). Additional sequencing methods such as DNA-Seq, Assay for Transposase-Accessible Chromatin sequencing (ATAC-Seq), and Chromatin Immunoprecipitation sequencing (ChIP-Seq) can identify functional regions affected by genetic variants. Gene editing techniques including Crispr-Cas9 that enable knockdown of genes or SNP-specific editing may be followed up by a readout on the effects of gene regulatory networks (115) (Table 6).

Conclusions

With large efforts such as TOPMed and the U.K. Biobank, in addition to specific disease studies, there is an ever-increasing amount of sequencing data available for studies of respiratory disease, critical care, and sleep medicine. Because of the complex nature of these studies, it is critical to include researchers with multiple backgrounds at the outset of study design, including clinician-scientists and epidemiologists who can enroll and phenotype subjects; laboratory personnel with skills in biobanking, sample management, and high-throughput sequencing; bioinformaticians, statisticians, and computational biologists who can manage and analyze data; and molecular biologists who can conduct functional validation studies. All of these experts must collaborate to design studies, interpret data, and present results. This should not discourage new investigators from participating in omics studies, as each person can provide complementary expertise. Specific recommendations regarding study design, analysis, and follow-up (Box 1) should serve as guides for starting a new sequencing study or for the critical appraisal of a completed study. Genomics is a rapidly evolving field, and researchers must keep abreast of best practices. However, general principles of study design and data reporting are likely to remain valid in the future. ■

This workshop report was prepared by a subcommittee of the ATS Section on Genetics and Genomics in the Assembly on Allergy, Immunology, and Inflammation.

Members of the subcommittee are as follows:

Craig P. Hersh, M.D., M.P.H.[1,2,3] (*Co-Chair*)
Gerard H. Koppelman, M.D., Ph.D.[4,5] (*Co-Chair*)
Ian M. Adcock, Ph.D.[6]
Juan C. Celedón, M.D., Dr.P.H.[7]
Michael H. Cho, M.D., M.P.H.[1,2,3]
David C. Christiani, M.D., M.P.H.[3,8,9,10]
Blanca E. Himes, Ph.D.[11]
Naftali Kaminski, M.D.[12]
Rasika A. Mathias, Sc.D.[13]
Deborah A. Meyers, Ph.D.[14]
John Quackenbush, Ph.D.[15]
Susan Redline, M.D., M.P.H.[3,16,17]
Katrina A. Steiling, M.D., M.Sc.[18]
Holly K. Tabor, Ph.D.[19]
Martin Tobin, Ph.D., M.B.Ch.B.[20,21]
Mark M. Wurfel, M.D., Ph.D.[22]
Ivana V. Yang, Ph.D.[23]

[1]Channing Division of Network Medicine, [2]Division of Pulmonary and Critical Care Medicine, and [16]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, Massachusetts; [3]Harvard Medical School, Boston, Massachusetts;

[4]Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, and [5]Groningen Research Institute for Asthma and COPD, University Medical Center Groningen, University of Groningen, Groningen, Netherlands; [6]Airways Disease Section, National Heart and Lung Institute, Imperial College London, London, United Kingdom; [7]Division of Pediatric Pulmonary Medicine, Allergy and Immunology, Children's Hospital of Pittsburgh of University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; [8]Department of Environmental Health, [9]Department of Epidemiology, and [15]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; [10]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts; [11]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania; [12]Pulmonary, Critical Care and Sleep Medicine, Yale University School of Medicine, New Haven, Connecticut; [13]Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University, Baltimore, Maryland; [14]University of Arizona College of Medicine, Tucson, Arizona; [17]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts; [18]Division of Computational Biomedicine, Boston University School of Medicine, Boston, Massachusetts; [19]Stanford Center for Biomedical Ethics, Department of Medicine, Stanford University, School of Medicine, Stanford, California; [20]Department of Health Sciences, University of Leicester, Leicester, United Kingdom; [21]National Institute for Health Research Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom; [22]Division of Pulmonary, Critical Care and Sleep Medicine, University of Washington, Seattle, Washington; [23]Department of Medicine, University of Colorado, Denver, Colorado.

Author Disclosures: C.P.H. served as a consultant for 23andMe, AstraZeneca, Concert Pharmaceuticals and Mylan; received research support from Boehringer Ingelheim and Novartis. I.M.A. received research support from Boehringer Ingelheim and Novartis. J.C.C. received research support from GlaxoSmithKline, Merck, and Pharmavite. M.H.C. received research support from GlaxoSmithKline. N.K. served as a consultant for Biogen, Boehringer Ingelheim, Indaloo, MMI, NuMedii, Samumed and Third Rock; on an advisory committee and has stock options with Moereae Matrix and Pliant; received research support from Miragen; holds patents on peripheral blood gene expression issue, biomarkers for assessing idiopathic pulmonary fibrosis, methods of treating or preventing fibrotic lung diseases with Aerosolized T3 with royalties paid by Lifemax, new therapies in pulmonary fibrosis with royalties paid by Biotech; has a pulmonary fibrosis intellectual property commercialized by Qyuitza. J.Q. was the owner and chairman of Genospace until January 2017. S.R. received research support from Beckman Coulter and Jazz Pharma. K.A.S. received author royalties from UpToDate for co-authoring chapter on gene expression; received research support from the LUNGevity Foundation; holds patent on biomarkers of COPD disease activity filed by Boston University (no royalties). M.D.T. received research support from GlaxoSmithKline and Pfizer. G.H.K. received research support from GlaxoSmithKline, the Lung Foundation of the Netherlands, TETRI Foundation, Teva, UBBO Emmius Foundation and Vertex Pharmaceuticals. D.C.C., B.E.H., R.A.M., D.A.M., H.K.T., M.M.W., and I.V.Y. reported no relationships with relevant commercial interests.

References

1 Campbell CD, Mohajeri K, Malig M, Hormozdiari F, Nelson B, Du G, *et al*. Whole-genome sequencing of individuals from a founder population identifies candidate genes for asthma. *PLoS One* 2014;9: e104396.
2 Kim WJ, Lim JH, Lee JS, Lee SD, Kim JH, Oh YM. Comprehensive analysis of transcriptome sequencing data in the lung tissues of COPD subjects. *Int J Genomics* 2015;2015:206937.
3 Smith D, Helgason H, Sulem P, Bjornsdottir US, Lim AC, Sveinbjornsson G, *et al*. A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. *PLoS Genet* 2017;13:e1006659.
4 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.
5 Qiao D, Lange C, Beaty TH, Crapo JD, Barnes KC, Bamshad M, *et al*.; Lung GO; NHLBI Exome Sequencing Project; COPDGene Investigators. Exome sequencing analysis in severe, early-onset chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2016;193:1353–1363.
6 Wain LV, Sayers I, Soler Artigas M, Portelli MA, Zeggini E, Obeidat M, *et al*. Whole exome re-sequencing implicates CCDC38 and cilia structure and function in resistance to smoking related airflow obstruction. *PLoS Genet* 2014;10:e1004314.
7 Petrovski S, Todd JL, Durheim MT, Wang Q, Chien JW, Kelly FL, *et al*. An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am J Respir Crit Care Med* 2017;196: 82–93.
8 Stuart BD, Choi J, Zaidi S, Xing C, Holohan B, Chen R, *et al*. Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat Genet* 2015;47: 512–517.
9 Caserta S, Kern F, Cohen J, Drage S, Newbury SF, Llewelyn MJ. Circulating plasma microRNAs can differentiate human sepsis and Systemic Inflammatory Response Syndrome (SIRS). *Sci Rep* 2016;6: 28006.
10 Degn M, Dauvilliers Y, Dreisig K, Lopez R, Pfister C, Pradervand S, *et al*. Rare missense mutations in P2RY11 in narcolepsy with cataplexy. *Brain* 2017;140:1657–1668.
11 Li J, Yu KH, Oehlert J, Jeliffe-Pawlowski LL, Gould JB, Stevenson DK, *et al*. Exome sequencing of neonatal blood spots and the identification of genes implicated in bronchopulmonary dysplasia. *Am J Respir Crit Care Med* 2015;192:589–596.
12 Hadinnapola C, Bleda M, Haimel M, Screaton N, Swift A, Dorfmuller P, *et al*.; NIHR BioResource-Rare Diseases Consortium, UK National Cohort Study of Idiopathic and Heritable PAH. Phenotypic characterization of EIF2AK4 mutation carriers in a large cohort of patients diagnosed clinically with pulmonary arterial hypertension. *Circulation* 2017;136:2022–2033.
13 Parker MM, Chase RP, Lamb A, Reyes A, Saferali A, Yun JH, *et al*. RNA sequencing identifies novel non-coding RNA and exon-specific effects associated with cigarette smoking. *BMC Med Genomics* 2017;10:58.
14 Shah R, Yeri A, Das A, Courtright-Lim A, Ziegler O, Gervino E, *et al*. Small RNA-seq during acute maximal exercise reveal RNAs involved in vascular inflammation and cardiometabolic health: brief report. *Am J Physiol Heart Circ Physiol* 2017;313: H1162–H1167.
15 Xu Y, Mizuno T, Sridharan A, Du Y, Guo M, Tang J, *et al*. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* 2016;1:e90558.
16 Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, *et al*. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;17:13.
17 Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, *et al*. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 2013;14:460–470.
18 Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, *et al*.; STrengthening the REporting of Genetic Association Studies. STrengthening the REporting of Genetic Association Studies

(STREGA): an extension of the STROBE statement. *PLoS Med* 2009; 6:e22.
19 Himes BE, Hunninghake GM, Baurley JW, Rafaels NM, Sleiman P, Strachan DP, *et al*. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *Am J Hum Genet* 2009;84: 581–593.
20 Nieuwenhuis MA, Vonk JM, Himes BE, Sarnowski C, Minelli C, Jarvis D, *et al*. PTTG1IP and MAML3, novel genomewide association study genes for severity of hyperresponsiveness in adult asthma. *Allergy* 2017;72:792–801.
21 Chen H, Cade BE, Gleason KJ, Bjonnes AC, Stilp AM, Sofer T, *et al*. Multiethnic meta-analysis identifies RAI1 as a possible obstructive sleep apnea-related quantitative trait locus in men. *Am J Respir Cell Mol Biol* 2018;58:391–401.
22 Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, *et al*. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* 2009;5: e1000429.
23 Hobbs BD, de Jong K, Lamontagne M, Bossé Y, Shrine N, Artigas MS, *et al*.; COPDGene Investigators; ECLIPSE Investigators; LifeLines Investigators; SPIROMICS Research Group; International COPD Genetics Network Investigators; UK BiLEVE Investigators; International COPD Genetics Consortium. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet* 2017;49:426–432.
24 Chapman KR, Tashkin DP, Pye DJ. Gender bias in the diagnosis of COPD. *Chest* 2001;119:1691–1695.
25 Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, *et al*.; eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genet Med* 2013;15:761–771.
26 Rogers AJ. Genome-wide association study in acute respiratory distress syndrome: finding the needle in the haystack to advance our understanding of acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2018;197:1373–1374.
27 Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;12:745–755.
28 Cho MH, McDonald ML, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, *et al*.; NETT Genetics, ICGN, ECLIPSE and COPDGene Investigators. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med* 2014;2: 214–225.
29 Bime C, Pouladi N, Sammani S, Batai K, Casanova N, Zhou T, *et al*. Genome-wide association study in african americans with acute respiratory distress syndrome identifies the selectin P ligand gene as a risk factor. *Am J Respir Crit Care Med* 2018;197:1421–1432.
30 Christie JD, Wurfel MM, Feng R, O'Keefe GE, Bradfield J, Ware LB, *et al*.; Trauma ALI SNP Consortium (TASC) investigators. Genome wide association identifies PPFIA1 as a candidate gene for acute lung injury risk following major trauma. *PLoS One* 2012;7:e28268.
31 Turnbull C. Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. *Ann Oncol* 2018;29:784–787.
32 Brody JA, Morrison AC, Bis JC, O'Connell JR, Brown MR, Huffman JE, *et al*.; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium; TOPMed Hematology and Hemostasis Working Group; CHARGE Analysis and Bioinformatics Working Group. Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* 2017; 49:1560–1563.
33 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, *et al*.; Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–291.
34 Collins R. What makes UK Biobank special? *Lancet* 2012;379: 1173–1174.
35 Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Soler Artigas M, *et al*.; UK Brain Expression Consortium (UKBEC); OxGSK Consortium. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* 2015;3: 769–781.
36 Chen KD, Chang PT, Ping YH, Lee HC, Yeh CW, Wang PN. Gene expression profiling of peripheral blood leukocytes identifies and validates ABCB1 as a novel biomarker for Alzheimer's disease. *Neurobiol Dis* 2011;43:698–705.
37 Celedón JC, Burchard EG, Schraufnagel D, Castillo-Salgado C, Schenker M, Balmes J, *et al*.; American Thoracic Society and the National Heart, Lung, and Blood Institute. An American Thoracic Society/National Heart, Lung, and Blood Institute Workshop Report: addressing respiratory health equality in the United States. *Ann Am Thorac Soc* 2017;14:814–826.
38 Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP, Colangelo L, *et al*. Genetic ancestry in lung-function predictions. *N Engl J Med* 2010;363:321–330.
39 Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, *et al*.; Mexico City Childhood Asthma Study (MCAAS); Children's Health Study (CHS) and HARBORS study; Genetics of Asthma in Latino Americans (GALA) Study, Study of Genes-Environment and Admixture in Latino Americans (GALA2) and Study of African Americans, Asthma, Genes & Environments (SAGE); Childhood Asthma Research and Education (CARE) Network; Childhood Asthma Management Program (CAMP); Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE); Genetic Research on Asthma in African Diaspora (GRAAD) Study. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet* 2011;43:887–892.
40 Manichaikul A, Hoffman EA, Smolonska J, Gao W, Cho MH, Baumhauer H, *et al*. Genome-wide study of percent emphysema on computed tomography in the general population: the Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association resource study. *Am J Respir Crit Care Med* 2014;189:408–418.
41 Cade BE, Chen H, Stilp AM, Gleason KJ, Sofer T, Ancoli-Israel S, *et al*. Genetic associations with obstructive sleep apnea traits in Hispanic/Latino Americans. *Am J Respir Crit Care Med* 2016;194:886–897.
42 Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, *et al*. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 2008;40:1092–1097.
43 Chung WH, Hung SI, Hong HS, Hsih MS, Yang LC, Ho HC, *et al*. Medical genetics: a marker for Stevens-Johnson syndrome. *Nature* 2004;428:486.
44 Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, *et al*.; CAAPA. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat Commun* 2016;7: 12522.
45 Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, *et al*. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med* 2016;375:655–665.
46 Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. *Nature* 2011;475:163–165.
47 Strouse JJ, Lobner K, Lanzkron S, Haywood C. NIH and National Foundation expenditures for sickle cell disease and cystic fibrosis are associated with PubMed publications and FDA approvals [abstract]. *Blood* 2013;122:1739.
48 Colby SL, Ortman JM. Projections of the size and composition of the US population: 2014 to 2060. Washington, D.C.: U.S. Census Bureau; 2015.
49 Paltoo DN, Rodriguez LL, Feolo M, Gillanders E, Ramos EM, Rutter JL, *et al*.; National Institutes of Health Genomic Data Sharing Governance Committees. Data use under the NIH GWAS data sharing policy and future directions. *Nat Genet* 2014;46: 934–938.
50 Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;39:1181–1186.
51 Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, *et al*. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med* 2017;19:249–255.

52 Goodman JL, Amendola LM, Horike-Pyne M, Trinidad SB, Fullerton SM, Burke W, *et al*. Discordance in selected designee for return of genomic findings in the event of participant death and estate executor. *Mol Genet Genomic Med* 2017;5:172–176.
53 Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, *et al*. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* 2012;44: 631–635.
54 Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, *et al*. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 2015;112:5473–5478.
55 Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, *et al*.; Centers for Mendelian Genomics. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 2015;97:199–215.
56 Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, *et al*. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010;42: 790–793.
57 MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, *et al*. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;508: 469–476.
58 Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, *et al*.; NHLBI GO Exome Sequencing Project. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI Exome Sequencing Project. *Am J Hum Genet* 2016; 99:791–801.
59 Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, *et al*.; Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* 2013;45:899–901.
60 MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, *et al*.; 1000 Genomes Project Consortium. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012;335:823–828.
61 Wang Z, Sadovnick AD, Traboulsee AL, Ross JP, Bernales CQ, Encarnacion M, *et al*. Nuclear receptor NR1H3 in familial multiple sclerosis. *Neuron* 2016;90:948–954.
62 Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 2012;109:1193–1198.
63 Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* 2011;8:469–477.
64 Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, *et al*. A beginner's guide to analysis of RNA sequencing data. *Am J Respir Cell Mol Biol* 2018;59:145–157.
65 Lei R, Ye K, Gu Z, Sun X. Diminishing returns in next-generation sequencing (NGS) transcriptome data. *Gene* 2015;557:82–87.
66 Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, *et al*. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics* 2011;12:S5.
67 Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res* 2011;21:2213–2223.
68 Liu Y, Ferguson JF, Xue C, Silverman IM, Gregory B, Reilly MP, *et al*. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLoS One* 2013;8:e66883.
69 Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* 2014;30:301–304.
70 Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11:31–46.
71 Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016; 17:333–351.
72 Raz T, Kapranov P, Lipson D, Letovsky S, Milos PM, Thompson JF. Protocol dependence of sequencing-based gene expression measurements. *PLoS One* 2011;6:e19287.
73 Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, *et al*. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* 2014;56: 61–64, 66, 68.
74 McCall MN, Illei PB, Halushka MK. Complex sources of variation in tissue expression data: analysis of the GTEx lung transcriptome. *Am J Hum Genet* 2016;99:624–635.
75 Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;14: 91.
76 Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* 2012; 99:248–256.
77 Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* 2012;13:484.
78 Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, *et al*. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013;14:R95.
79 SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 2014;32: 903–914.
80 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;29:365–371.
81 Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Med* 2010;2:84.
82 Brookes AJ, Robinson PN. Human genotype-phenotype databases: aims, challenges and opportunities. *Nat Rev Genet* 2015;16: 702–715.
83 Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, *et al*. Big data: astronomical or genomical? *PLoS Biol* 2015;13: e1002195.
84 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43: 491–498.
85 Himes BE, Koziol-White C, Johnson M, Nikolos C, Jester W, Klanderman B, *et al*. Vitamin D modulates expression of the airway smooth muscle transcriptome in fatal asthma. *PLoS One* 2015;10: e0134057.
86 Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;5:e1000477.
87 Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014;95: 5–23.
88 Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, *et al*.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;91:224–237.
89 Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, *et al*. A benchmark for RNA-seq quantification pipelines. *Genome Biol* 2016; 17:74.
90 Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, *et al*. Ensembl 2018. *Nucleic Acids Res* 2018;46:D754–D761.
91 O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–D745.
92 Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res* 2011;39:D19–D21.
93 Franks TJ, Colby TV, Travis WD, Tuder RM, Reynolds HY, Brody AR, *et al*. Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function. *Proc Am Thorac Soc* 2008;5:763–766.
94 Wansleeben C, Barkauskas CE, Rock JR, Hogan BL. Stem cells of the adult lung: their development and role in homeostasis, regeneration, and disease. *Wiley Interdiscip Rev Dev Biol* 2013;2:131–148.

95 Houseman EA, Kelsey KT, Wiencke JK, Marsit CJ. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics* 2015;16:95.
96 Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;18:220.
97 Seumois G, Chavez L, Gerasimova A, Lienhard M, Omran N, Kalinke L, *et al*. Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nat Immunol* 2014;15:777–788.
98 Dong X, Shi M, Lee M, Toro R, Gravina S, Han W, *et al*. Global, integrated analysis of methylomes and transcriptomes from laser capture microdissected bronchial and alveolar cells in human lung. *Epigenetics* 2018;13:264–274.
99 Prakadan SM, Shalek AK, Weitz DA. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat Rev Genet* 2017;18: 345–361.
100 Hu Y, An Q, Sheu K, Trejo B, Fan S, Guo Y. Single cell multi-omics technology: methodology and application. *Front Cell Dev Biol* 2018; 6:28.
101 Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. *Science* 2017;358:58–63.
102 Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;9:75.
103 Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, *et al*. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 2014;509:371–375.
104 Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, *et al*. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017;14:955–958.
105 Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, *et al*.; Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife* 2017;6:e27041.
106 Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* 2013;45:1238–1243.
107 Hao K, Bossé Y, Nickle DC, Paré PD, Postma DS, Laviolette M, *et al*. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* 2012;8:e1003029.
108 GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;45:580–585.
109 Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, *et al*.; GTEx Consortium. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;47:1091–1098.
110 Ardini-Poleske ME, Clark RF, Ansong C, Carson JP, Corley RA, Deutsch GH, *et al*.; LungMAP Consortium. LungMAP: the Molecular Atlas of Lung Development Program. *Am J Physiol Lung Cell Mol Physiol* 2017;313:L733–L740.
111 Cleary B, Cong L, Cheung A, Lander ES, Regev A. Efficient generation of transcriptomic profiles by random composite measurements. *Cell* 2017;171:1424–1436, e18.
112 Li CX, Wheelock CE, Sköld CM, Wheelock AM. Integration of multi-omics datasets enables molecular classification of COPD. *Eur Respir J* 2018;51:1701930.
113 Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;24:1248–1259.
114 Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. *Nat Rev Genet* 2017;18:425–440.
115 Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* 2014;32:347–355.
116 Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. *Am J Hum Genet* 2009;85:142–154.
117 Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 2014;30:2843–2851.
118 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
119 Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 2011;27:1691–1692.
120 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
121 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25: 2078–2079.
122 Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, *et al*. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 2018;201178
123 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, *et al*. Integrative genomics viewer. *Nat Biotechnol* 2011;29: 24–26.
124 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, *et al*. The human genome browser at UCSC. *Genome Res* 2002;12: 996–1006.
125 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
126 Conomos MP, Gogarten SM, Brown L, Chen H, Rice K, Sofer T, *et al*. GENESIS: GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. 2018. R package version 2.12.0, https://github.com/UW-GAC/GENESIS.
127 Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, *et al*. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;47:284–290.
128 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.
129 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29: 15–21.
130 Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 2012;7: 562–578.
131 Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* 2013;10: 71–73.
132 Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;31:166–169.
133 Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525–527.
134 Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;12:323.
135 Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;28:2184–2185.
136 Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, *et al*. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* 2015;43:e140.
137 Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 2010;26:136–138.
138 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15: 550.
139 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–140.
140 Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;15:R29.
141 Burden CJ, Qureshi SE, Wilson SR. Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ* 2014;2: e576.
142 Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* 2017;14:687–690.

Personalis EX2019

143 Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013;31:46–53.
144 Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;22:2008–2017.
145 Wang W, Qin Z, Feng Z, Wang X, Zhang X. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* 2013; 518:164–170.
146 Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;7:1009–1015.
147 Shi Y, Jiang H. rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. *PLoS One* 2013;8:e79448.
148 Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, *et al*. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* 2018;50:151–158.
149 Rogé X, Zhang X. RNAseqViewer: visualization tool for RNA-Seq data. *Bioinformatics* 2014;30:891–892.
150 Wu E, Nance T, Montgomery SB. SplicePlot: a utility for visualizing splicing quantitative trait loci. *Bioinformatics* 2014;30:1025–1026.
151 Ryan MC, Cleland J, Kim R, Wong WC, Weinstein JN. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics* 2012;28:2385–2387.
152 Liu Q, Chen C, Shen E, Zhao F, Sun Z, Wu J. Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer. *Genomics* 2012;99:178–182.
153 Li Y, Oosting M, Deelen P, Ricaño-Ponce I, Smeekens S, Jaeger M, *et al*. Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nat Med* 2016;22:952–960.
154 Ferreira MA, Jansen R, Willemsen G, Penninx B, Bain LM, Vicente CT, *et al*.; Australian Asthma Genetics Consortium Collaborators. Gene-based analysis of regulatory variants identifies 4 putative novel asthma risk genes related to nucleotide synthesis and signaling. *J Allergy Clin Immunol* 2017;139:1148–1157.
155 Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, *et al*. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 2016;167:1853–1866, e17.