Integrating Genetics and Omics to Understand Chronic Obstructive Pulmonary Disease

Edwin K. Silverman, MD, PhD and Brian D. Hobbs, MD, MMSc

Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

ABSTRACT

The availability of comprehensive assessments of specific types of biologically active molecules, referred to as Omics, in large study populations provides unique opportunities for the discovery of new pathobiological mechanisms for complex diseases like chronic obstructive pulmonary disease (COPD). Omics data can assist in identifying key genes that are driving genetic associations to COPD. However, despite substantial progress in delineating the genetic determinants of COPD, the biological networks influencing COPD remain largely undefined. Multiple methods have recently been developed to integrate multiple Omics data and identify key biological networks. These methods are already beginning to provide new insights into COPD pathogenesis and heterogeneity. **(BRN Rev. 2020;6(2):104-17)**

Corresponding author: Edwin K. Silverman, ed.silverman@channing.harvard.edu

Key words: Chronic obstructive pulmonary disease. Genetic association. Network medicine. Omics.

Correspondence to:
Dr. Edwin K. Silverman
Channing Division of Network Medicine,
Brigham and Women's Hospital,
181 Longwood Avenue
Boston, Massachusetts 02115, USA
E-mail: ed.silverman@channing.harvard.edu

Received in original form: 12-06-2020
Accepted in final form: 02-09-2020
DOI: 10.23866/BRNRev:2019-0008

I. INTRODUCTION

Like most diseases that are major public health problems, chronic obstructive pulmonary disease (COPD) is a complex syndrome influenced by multiple genetic and environmental determinants acting within a development context[1]. Traditionally, investigators trying to understand the biological mechanisms for a complex disease have studied selected molecules that they hypothesized would influence disease pathogenesis based on scientific intuition and/or analogies from other disease mechanisms. The development of large-scale biological assessments, referred to as "Omics", allows for comprehensive investigation of particular biological entities rather than selection of a small number of potentially important candidate molecules for study[2]. For example, instead of selecting a candidate gene for a genetic association study, all of the genes in the human genome can be assessed for association to a complex disease. The potential to gain additional pathobiological insights by integrating the information provided by multiple Omics data types is substantial and will be the focus of this review. Our goal is to provide an overview of Omics data types and analytical approaches, using applications to COPD as examples, but not to provide an exhaustive review of every multiple Omics integration method that has been proposed.

II. OVERVIEW OF OMICS DATA TYPES

The major molecular Omics data types are shown in figure 1. Although most of the three billion nucleotide base pairs of DNA sequence in the human genome are identical in all people, there are about 10 million common genetic variants, referred to as single nucleotide polymorphisms (SNPs), and many more rare genetic variants. These genetic variants can be readily assessed using specific assays for single genetic variants, genome-wide SNP panels that provide reasonable coverage of common genetic variation, and whole genome sequencing for comprehensive genetic variation assessment. Epigenetic alterations to the DNA sequence, such as methylation and histone acetylation, can influence gene regulation. Commercially available panels of DNA methylation marks can be assessed for hundreds of thousands of genomic locations; DNA sequencing before and after bisulfite conversion can provide comprehensive assessment of DNA methylation marks. High throughput assays for histone acetylation have been more challenging to develop.

Messenger RNA (mRNA) previously was assessed with microarrays but currently is analyzed with RNA-sequencing. Proteins can be measured as single analytes (e.g., enzyme-linked immunosorbent assays, ELISA), panels of selected proteins (e.g., Luminex, OLINK, SomaLogic), or using large-scale mass spectrometry approaches. Metabolites of various classes can be measured with targeted panels or untargeted assays; the identification of novel metabolites from untargeted assays can be quite challenging. Other Omics data types, which we will not focus on in this review, include microRNAs and other small RNAs as well as the microbiome.

The quality control and analytical approaches differ for each of these Omics data types, but there are four fundamental study design components for all Omics studies. First, a study population must be selected, which can be cases and controls, general population samples, or family-based units. Second, the phenotype of interest

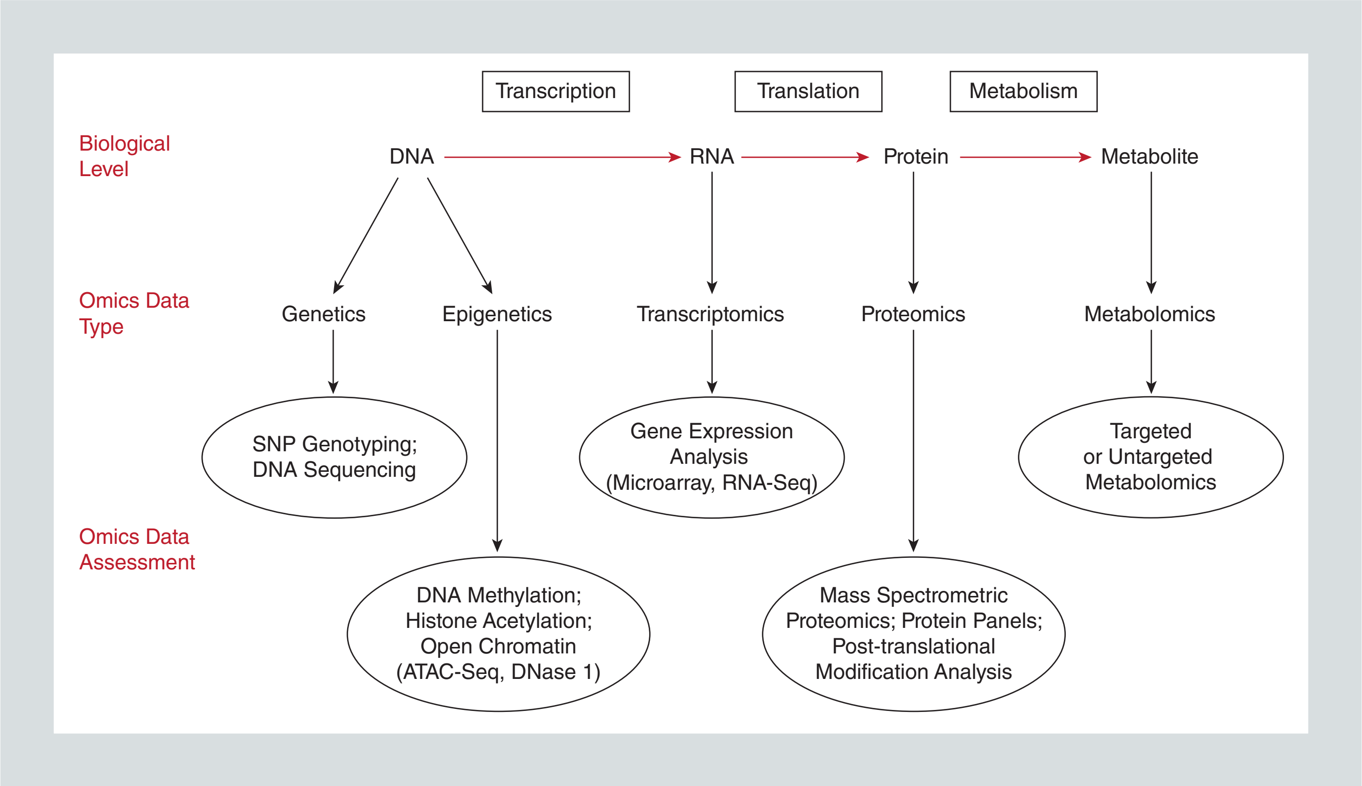
Transcription
Translation
Metabolism
Biological Level
DNA
RNA
Protein
Metabolite
Omics Data Type
Genetics
Epigenetics
Transcriptomics
Proteomics
Metabolomics
SNP Genotyping; DNA Sequencing
Gene Expression Analysis (Microarray, RNA-Seq)
Targeted or Untargeted Metabolomics
Omics Data Assessment
DNA Methylation; Histone Acetylation; Open Chromatin (ATAC-Seq, DNase 1)
Mass Spectrometric Proteomics; Protein Panels; Post-translational Modification Analysis

FIGURE 1. **Biological levels and Omics:** A simplified representation of key biological processes and related Omics data types is shown. DNA is transcribed to RNA, and mRNA is translated to proteins, which can then undergo post-translational modifications. Proteins can function as enzymes to catalyze reactions between different types of metabolites; metabolites also serve as building blocks for other key molecules in the cell *(modified from Network Medicine: Complex Systems in Human Disease and Therapeutics, Chapter 1, Loscalzo/Barabasi/Silverman[2], with permission)*.
mRNA: Messenger RNA; SNP: single nucleotide polymorphism.

needs to be determined; both categorical phenotypes (e.g., disease affection status) and quantitative phenotypes (e.g., quantitative computed tomography [CT] emphysema) can be used. Third, the laboratory assessment method for the Omics data type needs to be selected. Finally, the analytical approaches to adjust for quality control issues (e.g., batch effects, outlier removal, normalization, imputation of missing values) and test for relationships of disease phenotypes to Omics analytes need to be performed.

Key decisions in Omics studies include the biological sample to be analyzed. Increasingly, single cell assays are utilized, although single cell Omics data suffer from sparsity due to both biologic and technical reasons, making data analysis challenging[3, 4]. Single cell types (many cells of the same type, rather than individual cells) or tissue samples have been widely used for Omics analysis. Important study population-related issues include the stage of disease (mild versus severe) of included subjects and the sample size for analysis.

Although it is beyond the scope of this review, disease-related phenotypes are often considered as an additional key data type in Omics analysis, particularly given the high dimensionality of CT and other imaging data in COPD and other complex diseases. Phenotype specificity has implications for discovery of Omics associations with COPD. Due to the misclassification bias inherent to COPD defined by International Classification of Disease (ICD) codes or self-report, these COPD definitions resulted in reduced discovery of genetic risk loci compared to a spirometric definition of COPD in the UK Biobank[5]. High quality "deep" phenotyping is germane to assuring molecular Omics studies give useful biologic insights[6, 7], particularly with complex diseases that are as phenotypically heterogeneous as COPD. Two of the integrative Omics methods that we will discuss in this review - "integrative phenotyping framework" and sparse multiple canonical correlation network analysis (SmCCNet) - were designed to incorporate phenotype information into the analysis of multiple Omics data.

III. GENETICS OF COPD

As we have recently reviewed, there is strong evidence for genetic influences on COPD susceptibility and on COPD-related traits[8]. Rare Mendelian syndromes, such as alpha-1 antitrypsin (AAT) deficiency and cutis laxa, can include COPD as part of their phenotypic manifestations[9, 10]. Although COPD unrelated to AAT deficiency is strongly influenced by cigarette smoking, genome-wide association studies (GWAS) in large numbers of COPD cases and control subjects have identified 82 genomic regions with strong statistical evidence for association to COPD (Fig. 2)[11].

Despite this substantial progress in COPD genetics, there are a number of challenges involved in translating these genetic associations to new pathobiological insights. First, the associations highlight genomic regions of interest, but they do not provide adequate resolution to determine the key functional variants driving the statistical association between phenotype and genotype[12]. Second, even if the key genetic variant or variants in a GWAS region can be found, additional work is needed to determine which gene is influenced by those functional variants—it is often not the closest gene[13]. Third, the effect sizes of the many individual COPD GWAS loci are quite modest, suggesting that COPD is a polygenic condition. Utilizing genome-wide genetic variation in a polygenic risk score improves predictive ability for COPD[14]. However, to provide pathobiological insights and to dissect COPD heterogeneity, the specific genes involved in genetic association regions need to be identified. Some progress has been made, including the discovery of functional genetic variation influencing *HHIP*[15], *FAM13A*[16], and *TGFB2*[17], but much more work is needed. Omics data can assist in this challenging process.

IV. INTEGRATING GENETICS AND A SINGLE OMICS TYPE

Different Omics data types may be more relevant for different subtypes (subsets of COPD subjects with a shared pathobiological mechanism) of the heterogeneous COPD syndrome. A major limitation is that despite extensive efforts with unsupervised clustering analysis[18], visual and quantitative CT analysis[19], and transcriptomic-based clustering efforts[20], there is still not a consensus regarding COPD subtypes. One generally accepted distinction in COPD is that subjects with AAT deficiency have unique clinical and biological features consistent with a COPD subtype. As shown in table 1, the key

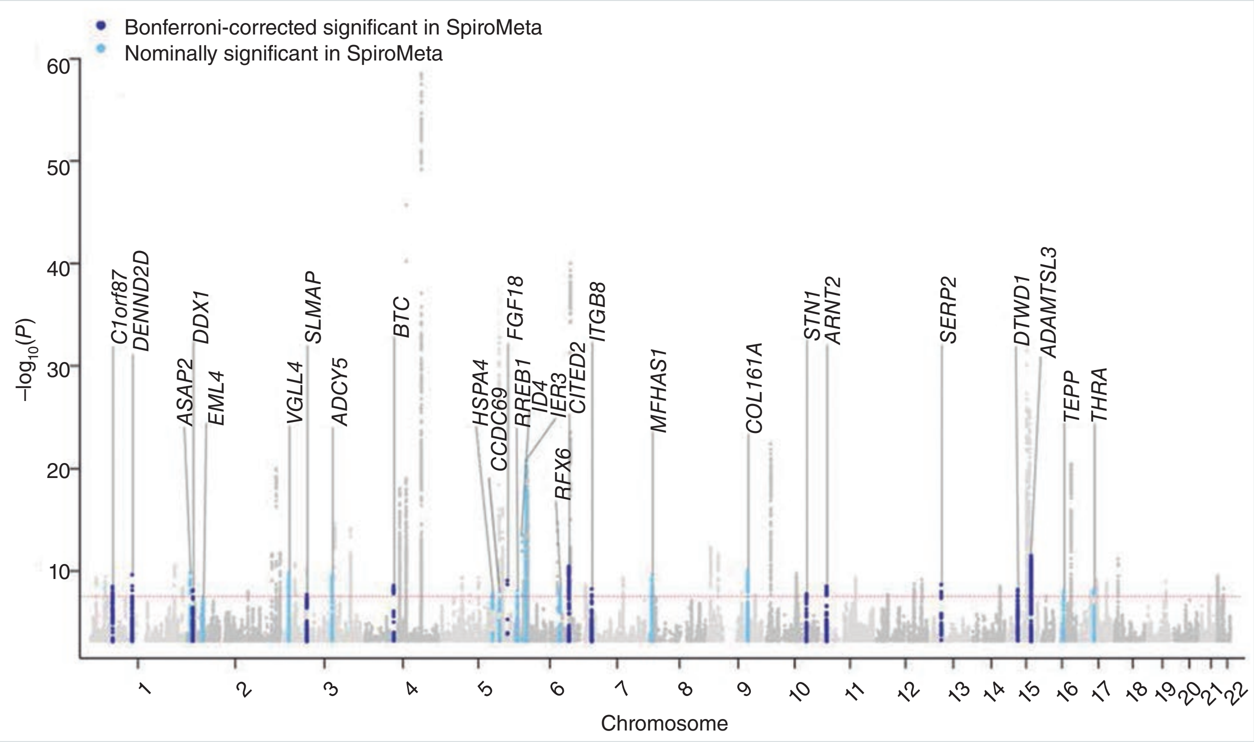

FIGURE 2. **International COPD Genetics Consortium and UK Biobank GWAS for COPD.** Manhattan plot demonstrating 82 genome-wide significant associations to COPD. Novel associations (not previously reported for COPD or lung function) are labeled with the nearest gene, and replication in the SpiroMeta cohort for lung function phenotypes is indicated *(from Sakornsakolpat et al.[11], Nature Genetics 2019, with permission)*.

COPD: chronic obstructive pulmonary disease; GWAS: genome-wide association studies.

TABLE 1. Omics in COPD With and Without Alpha-1 Antitrypsin (AAT) deficiency

	Genetics	Epigenetics	Transcriptomics	Proteomics	Metabolomics
AAT Deficiency	Missense SNP of major effect	Unknown	Non-diagnostic	Severely reduced plasma AAT level	Unknown
COPD without AAT Deficiency	Polygenic effects	Likely important, but key sites unclear	Some expression differences noted (? Cause vs. Effect)	Some protein level differences (? Cause vs. Effect)	Some metabolomic differences (? Cause vs. Effect)

COPD: chronic obstructive pulmonary disease.

biological mechanism for AATD is a missense variant in the *SERPINA1* gene sequence, which leads to a single amino acid change in the protein sequence and reduced circulating AAT protein levels[9]; thus, genetics and proteomics are key Omics data types for AATD. For non-AATD COPD, the roles of different Omics data types are unclear. Despite some evidence for associations of epigenetic marks, transcriptomics, proteomics, and metabolomics with COPD, consistently replicated associations have been challenging to find, and it is uncertain which Omics differences are causes or effects of the pathological changes in COPD without AATD. The issue of whether these non-genetic Omics measurements are markers of pathologic change or the cause of observed changes is not unique to COPD and is a problem across many complex diseases.

For each of these Omics data types, genetic determinants can be identified: expression quantitative trait loci (eQTLs) for transcriptomics[21], protein QTLs (pQTLs) for proteomics[22], methylation QTLs (mQTLs) for epigenetics[23], and metabolite QTLs for metabolomics[24]. When the genetic determinants of these Omics data types are located near a coding gene, the term "cis" QTLs is used. Almost every gene will have cis-QTLs for related Omics data types (detectable if a large enough sample size is used)[25,26], so finding a QTL near a GWAS functional variant is not proof that variant influences the gene related to that QTL. Statistical co-localization methods have been developed to determine statistically whether a genomic region associated with both a disease phenotype and an Omics QTL is likely to be driven by the same association signal[27]; however, multiple independent QTLs exist in many genetic regions[28]; which violates the assumptions of some colocalization methods.

Long-range genetic effects on Omics data types, known as trans-QTLs, have been more challenging to discover. Larger sample sizes are typically required to find trans than cis-QTL effects, but trans relationships can help to identify long-range biological influences. For example, Sun and colleagues[29] built a Plasma Proteome Genomic Atlas based on SomaLogic assays for 3622 proteins in 3301 participants; they found 1104 trans-pQTL associations, some of which led to hypotheses about biological mechanisms, by leveraging molecular pathways, protein-protein interactions (PPI), variant annotation, eQTLs, and chromatin interactions.

Lamontagne and colleagues[30] integrated GWAS results from the International COPD Genetics Consortium (ICGC) with the Laval/UBC/Groningen lung tissue eQTL database to interrogate likely causal genes within COPD and lung function GWAS loci. They utilized four approaches: 1) Bayesian colocalization, to determine if there was statistical evidence that the GWAS and lung eQTL associations were driven by the same signal; 2) transcriptome-wide association study (TWAS), to test for association of COPD to the estimated lung gene expression levels in the ICGC based on the genetic component of gene expression determined in the lung gene expression dataset; 3) Mendelian Randomization, to determine if using lung eQTL SNPs as instrumental variables can predict COPD; and 4) S-PrediXcan, an alternative approach to estimate gene expression from its genetic components by using an ElasticNet Model. For three COPD/lung function GWAS loci, all four methods implicated the same gene (*DSP*, *CIGALT1*, and *THSD4*). In 60 of the 129 GWAS loci examined, a top candidate gene for the GWAS signal was identified. Sakornsakolpat and colleagues[31] also used S-PrediXcan to estimate gene expression based on genome-wide association data for severe COPD and quantitative emphysema; they implicated key genes within several GWAS loci and found novel associations as well. These predictive approaches are valuable, but they have limitations. For example, correlation between adjacent SNPs (known as linkage disequilibrium) can affect the estimated genetic influences on gene expression using approaches like TWAS and S-PrediXcan. Thus, investigations that directly assess molecular interactions of putative functional variants with GWAS genes (e.g., chromatin conformation capture[15]) and that measure the impact of the genetic variant on gene expression (e.g., luciferase reporter assays) remain essential to confirm key genes within GWAS loci.

QTL analysis of plasma from COPD cases and controls by Sun and colleagues[22] led to the identification of many pQTLs. This work clarified the previously confusing relationship of the COPD GWAS SNP near *AGER*, which encodes the sRAGE protein biomarker—likely the most robust protein biomarker for COPD identified thus far[32]. Since the directions of association for that top *AGER* SNP, rs2070600, with COPD and with sRAGE protein levels were opposite, including the pQTL SNP genotype and sRAGE levels together in a predictive model provided stronger association with COPD.

Morrow and colleagues[23] performed methylation QTL analysis in lung tissue samples from COPD cases and controls, and they found that COPD GWAS loci were enriched with mQTLs. Using statistical colocalization, they identified several genomic regions, including *EEFSEC* and *IL27*, in which the mQTL and COPD GWAS signal appeared to colocalize.

V. INTEGRATING MULTIPLE OMICS: RATIONALE

There are many reasons why integrating different Omics data types should be pursued (Table 2). First, Omics data have substantial variability unrelated to the disease process of interest. This Omics "noise" can be technical - related to batch effects and measurement error, and biological -related to sex, age, environmental exposures, etc. Even if two different Omics data types are capturing the same disease-related biological signals, their technical and biological variability may be unrelated to each other; thus, true disease-related biological signals may be identified by implicating shared molecules and pathways in different Omics data types. Second, integrating Omics data types could provide insights into biological mechanisms for genetic variation. As discussed previously, quantitative trait loci between functional genetic variants and Omics data can provide insight into gene regulatory mechanisms. Third, different Omics data types can reflect different time scales for biological processes. Except for rare somatic mutation events, the DNA sequence of an individual is invariant over their lifetime. Precise measurements of the half-lives of Omics data types have been limited. Gene expression levels can change rapidly in response to environmental changes[33], since the half-lives of mRNAs are typically less than 20 minutes. The proteins that they encode can have widely variable lifespans; lung elastin likely lasts an entire lifetime[34], but other proteins have half-lives less than one day[35]. The metabolites produced by those proteins have relatively short life spans, and they are often assessed by flux balance analysis[36]. Epigenetic marks are typically longer lasting but can change in response to environmental exposures[37]. By reflecting different time scales, different Omics data types have the potential to give alternate perspectives

TABLE 2. Rationale for integrating multiple Omics data types in complex disease

Challenge	Rationale	Examples/Comments
Measurement error	Reduce noise from a single Omics data type; accentuate correlated signal across multiple data types	Different technical artifacts for different Omics types
Uncertain pathobiological mechanisms	Understand biological mechanisms for genetic variation	Similar to quantitative trait loci (QTLs) for individual Omics data types
Single Omics data may not capture relevant signals	Different time scales are captured	Genetics—lifetime; RNA—minutes; Proteins—hours to years; Epigenetics—long
Biological levels do not work in isolation	Interactions between biological levels can be found	miRNA regulating mRNA; methylation regulating mRNA; mRNA translated to proteins; protein enzymes control metabolite production
Complex diseases do not act at a single biological level	Generate more accurate biological models of disease	Feed-back and Feed-forward regulation

mRNA: messenger RNA; miRNA: microRNA.

on disease pathobiology. Fourth, assessment of multiple Omics data can allow identification of interactions between biological levels. For example, DNA methylation marks can regulate gene expression levels, and proteins (e.g., enzymes) can control metabolite production. Finally, and perhaps most importantly, biological systems operate on multiple biological levels. Integrating multiple Omics data types across these levels could provide more accurate biological models, with regulatory motifs such as feed-back and feed-forward regulation.

The recognition that biological systems function as molecular networks has helped to crystallize thinking about combining multiple Omics data in the field of Network Medicine[2,38]. The various methods utilized to build molecular networks have been recently reviewed by members of the International Network Medicine Consortium[39]. Although genetic studies of complex diseases have traditionally focused on analyses of single variants or clustered groups of variants, the role of biological networks has received increased attention through the Omnigenic Model of complex diseases[40]. As shown in figure 3, central goals of Network Medicine approaches to complex diseases include understanding gene regulation (since most complex disease genetic variants are regulatory and do not alter the protein-coding sequence), determining biological function (since most complex disease genetic variants have previously unknown biological effects), and defining disease pathobiology (which is necessary for improved diagnosis and treatment). "Bottom-up" approaches begin with the identification of functional genetic variants and their relationship to key disease genes, often with evidence provided from Omics data. "Top-down" approaches begin with large-scale Omics data sets and then use various types of network models, including correlation-based networks, gene regulatory networks, and protein-protein interaction networks. Both Bottom-up and Top-down approaches aspire to address one or more of the central goals of Network Medicine.

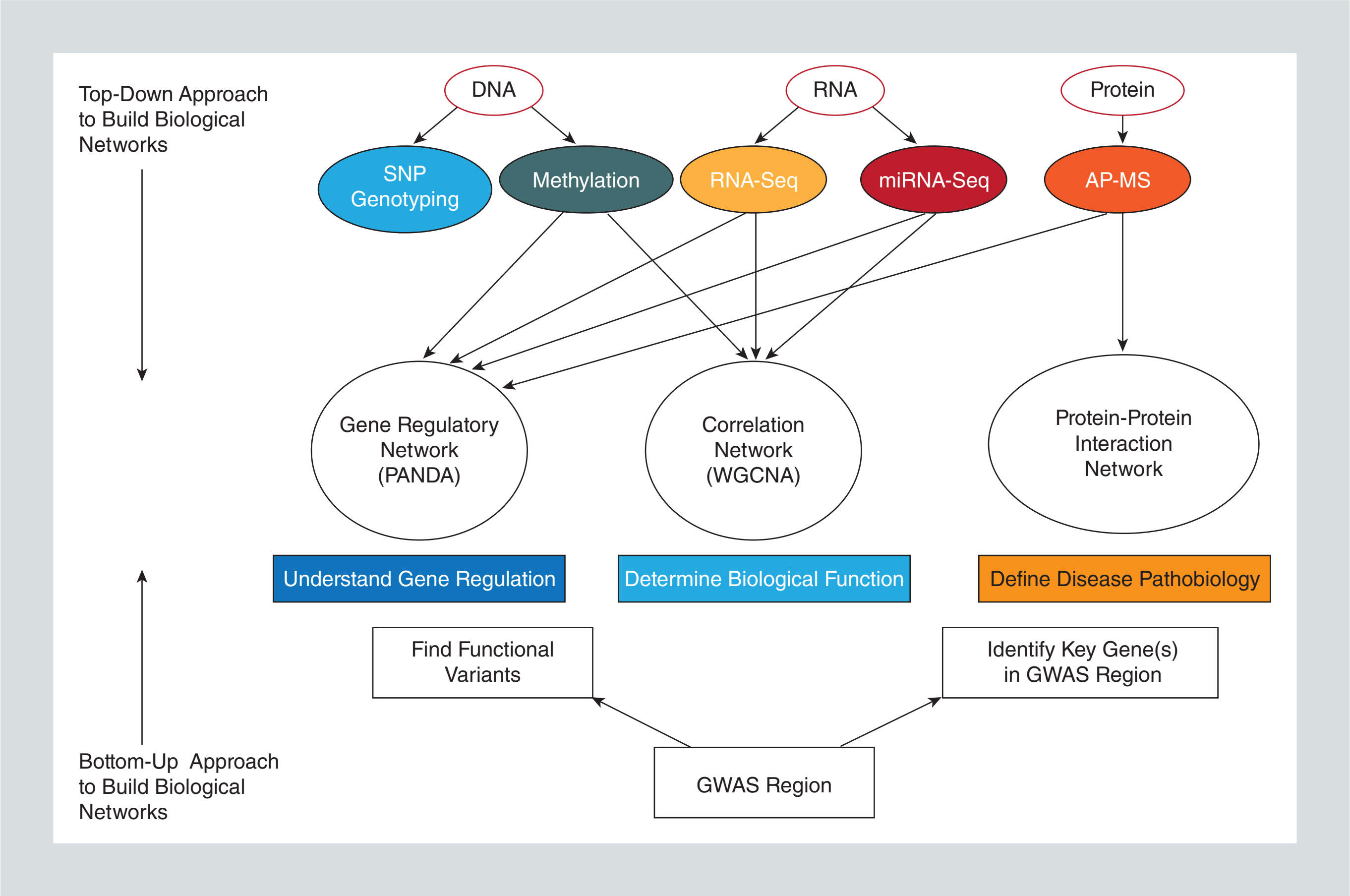
Top-Down Approach to Build Biological Networks
DNA
RNA
Protein
SNP Genotyping
Methylation
RNA-Seq
miRNA-Seq
AP-MS
Gene Regulatory Network (PANDA)
Correlation Network (WGCNA)
Protein-Protein Interaction Network
Understand Gene Regulation
Determine Biological Function
Define Disease Pathobiology
Find Functional Variants
Identify Key Gene(s) in GWAS Region
GWAS Region
Bottom-Up Approach to Build Biological Networks

FIGURE 3. **Comparison of Top-Down and Bottom-Up Approaches to Build Biological Networks.** Biological networks can be built from the bottom-up, starting with GWAS regions and identifying the key genes and the functional variants that impact those genes. Networks can be extended from those genes using tools that assess binding with other proteins (e.g., tandem affinity purification, co-immunoprecipitation) as well as hypothesis-based molecular biology experiments. Top-down approaches begin with Big Data assessments of key biological molecules like DNA (with SNP genotyping based on commercial panels or sequencing), RNA (with RNA-seq), proteins (with single analytes, commercial panels, or affinity purification/mass spectrometry [AP-MS]), and metabolites (with targeted or untargeted assays). Various types of networks can be built, including correlation-based networks, gene regulatory networks, and protein-protein interaction networks. Ultimately, bottom-up and top-down approaches may converge to give insights into gene regulation, biological function, and disease pathobiology relevant to COPD *(from EK Silverman[8], Annual Review of Physiology 2020, with permission)*.
COPD: chronic obstructive pulmonary disease; GWAS: genome-wide association studies; SNP: single nucleotide polymorphism; WGCNA: weighted gene correlation network analysis.

VI. INTEGRATING MULTIPLE OMICS: METHODS

Numerous methods have been proposed to integrate multiple Omics data, which have been recently reviewed[41-45]. Key goals of multiple Omics integration include dissecting disease heterogeneity, understanding disease pathogenesis, and predicting disease outcomes (Table 3).

Many of these integrative methods attempt to combine Omics data agnostically in an effort to reduce biological noise and to use multiple Omics to reclassify a complex disease into subtypes. For example, similarity network

Table 3. Applications of multiple Omics data in COPD

Goals of integrating multiple Omics data	Examples of methods to integrate multiple Omics	Examples of multiple Omics applications to COPD
Dissect disease heterogeneity	Similarity network fusion; Integrative phenotyping framework	Li (2018)[57]; Kim (2015)[54]
Understand disease pathogenesis	Weighted gene correlation network analysis; Sparse multiple canonical correlation network analysis	Mastej (2020)[58]; Shi (2019)[50]
Predict disease outcomes	Supervised machine learning (e.g., random forests) and deep learning (e.g., convolutional neural networks)	Not yet reported

COPD: chronic obstructive pulmonary disease.

fusion (SNF) generates similarity networks among individuals for a particular Omics data type, and then "fuses" the similarity networks for each Omics data type into an overall network model that leverages shared and unique information from each data type[46]. By creating a separate subject-to-subject similarity network for each Omics data type, the impact of each Omics type on the overall similarity matrix is not influenced by the number of features within an Omics data type. Subjects are clustered into subtypes based on their relationship to other subjects within this fused similarity network. Entropy-based consensus clustering (ECC) is another related method for agnostically combining Omics data[47]; multiple Omics data can be utilized to provide optimal molecular partitioning of a study population into discrete clusters, which potentially represent distinct molecular subtypes of a complex disease.

Network models have also been utilized to provide insights into disease pathobiology. Correlation-based networks, such as weighted gene correlation network analysis (WGCNA), have been widely used for analyses of single Omics data[48]. WGCNA network modules based on gene expression analysis have been compared between lung tissue, induced sputum, and peripheral blood in COPD[49]. Shi and Kechris[50] developed another unsupervised correlation-based network approach, parse Multiple Canonical Correlation Network Analysis (SmCCNet), which analyzes multiple Omics data to identify networks related to a quantitative disease-related phenotype of interest. They applied their approach to two Omics data types measured in blood (miRNA and mRNA) in small numbers of COPD cases and controls, with quantitative phenotypes of forced expiratory volume in one second (FEV_1) and CT quantitative emphysema. Other Network Medicine approaches leverage the known biological relationships between different Omics data types to provide insight into disease pathogenesis. For example, the Passing Attributes between Networks for Data Assimilation (PANDA) approach creates gene regulatory networks related to a disease of interest by using gene expression data, protein-protein interaction networks, and transcription factor binding site information in a message passing approach[51]. Although PANDA applications thus far have focused on transcriptomic data, other Omics data could be incorporated within this framework in the future.

Multiple Omics data can also be used to predict disease-related outcomes. Supervised machine learning approaches, including both traditional machine learning (e.g., random forests) and deep learning (e.g., convolutional neural networks) can be applied to multiple Omics data in order to make disease-related predictions[41]. Relevant COPD-related outcomes could include COPD exacerbations, disease progression, and mortality. The challenge of having more biological features than clinical observations increases the risk of overfitting machine learning models based on multiple Omics data; use of separate training and test data sets, optimally with inclusion of a completely separate validation dataset, should be strongly considered.

VII. INTEGRATING MULTIPLE OMICS: APPLICATIONS IN COPD

Although we are at the beginning of the multiple Omics era in complex diseases, there have already been several scientific efforts to combine multiple Omics data to understand COPD heterogeneity and COPD pathogenesis. Machine learning approaches have been applied to clinical and imaging data in COPD subjects to predict outcomes such as mortality[52] and to single Omics data, including blood miRNA to predict lung cancer risk in COPD patients[53]. However, multiple Omics data have not yet been reported for machine learning disease prediction in COPD.

Kim, Kaminski and colleagues[54] developed a model-free integrative Omics approach, which they named an "integrative phenotyping framework." After the data are pre-processed to remove batch effects, features are combined across Omics data types, dimension reduction is performed, and feature intensities are smoothed. Subsequently, clustering is done to identify disease subtypes. This approach was applied to a data set that included clinical features, miRNA, and mRNA from lung tissue samples of 319 COPD and interstitial lung disease (ILD) subjects—which was divided into a training and testing data set. In addition to finding clusters of subjects that were predominantly COPD or ILD, a mixed cluster of subjects with COPD-like molecular features who had been diagnosed as ILD was found. The integrative phenotyping framework has many advantages; one of its limitations is that Omics data can only be combined in a pairwise fashion.

Kusko, Kaminski and colleagues[55] utilized lung tissue samples from COPD (both with and without emphysema) and idiopathic pulmonary fibrosis (IPF) patients for transcriptomics analysis. They generated miRNA and mRNA transcriptomic data from lung tissue, and they used miRNA binding predictions and functionally validated miRNA/mRNA regulatory relationships to build miRNA/mRNA gene regulatory networks using the MirConnX software[56]. Networks were built using transcripts that were differentially expressed: 1) in the same direction in both IPF and emphysema, 2) when comparing emphysema versus control, and 3) when comparing IPF versus control lung tissue samples. All three approaches implicated miR96 as a potentially key miRNA.

Li and Wheelock[57] used similarity network fusion to classify individuals with mild-moderate COPD from nonsmoking controls and smokers with normal spirometry in a sample of 52 female subjects from the COSMIC study. They used mRNA, miRNA, proteomics, and metabolomics data from different biospecimens (bronchoalveolar lavage [BAL] fluid, BAL cells, bronchial epithelial cells, and serum). A total of nine Omics/biospecimen combinations were included. Permutation testing was used to determine whether improvements in COPD classification accuracy were statistically significant. Although the specific Omics/sample sets included impacted the classification rate, prediction accuracy generally increased with larger numbers of Omics data sets from one to seven that were included. Replication of these findings in additional cohorts will be necessary, along with further assessments to ensure that over-fitting the large amount of Omics data with a small number of subjects was not problematic.

Multiple Omics data can also be used to understand COPD pathogenesis. Recently, Mastej and Kechris[58] used SmCCNet with two Omics data types (metabolomics and proteomics of plasma) and quantitative COPD-related phenotypes (FEV_1 and CT quantitative emphysema) in 1008 COPDGene study participants. A network of seven metabolites and thirteen proteins was significantly correlated to lung function, and a network of ten metabolites and thirteen proteins was significantly correlated to emphysema. Interestingly, only two proteins (Troponin T and Hemojuvelin) and none of the metabolites overlapped between these two networks.

VIII. KEY CHALLENGES FOR APPLYING MULTIPLE OMICS IN COPD

Leveraging Omics data effectively will require both larger data sets in which multiple Omics data are measured in relevant biospecimens within the same subjects and new analytical methods to integrate these data types more effectively. The integration of common genetic variants with single Omics data types has become standard; additional methodological research will be required for rare genetic determinants, since those analyses will be underpowered with standard genetic association approaches. Rare variant association analysis approaches like linear mixed models in SAIGE hold promise[59]. Larger data sets are needed from biologically relevant specimen types; this limitation is being overcome with large projects like the NHLBI Trans-Omics for Precision Medicine program (TOPMed).

Further research is needed to validate molecular network models developed with multiple Omics data using both cell-based and animal models. Clarification is needed regarding what constitutes appropriate validation of a network. Is seeing a consistent signal after CRISPR-Cas9 manipulation of a genetic variant in a transformed cell line adequate? Are primary cells required? Where do animal models fit in? Further consideration of appropriate biological read-outs is required. Moreover, validation of an entire, complex molecular network is unrealistic; typically, a subset of key network relationships is chosen for investigation.

Finally, new network analysis methods are needed to integrate Omics data types across multiple biological layers in ways that are pathobiologically informative. Multi-level network models are easy to draw schematically but challenging to analyze rigorously. The development of models that incorporate the dynamics of biological systems and their application to longitudinal multiple Omics data, as recently reported in pre-diabetes[60], will be essential.

IX. CONCLUSION

The creation of large-scale Omics data sets in well-characterized human populations has the potential to transform studies of COPD pathobiology. Although assessment of single Omics associations with COPD is important, integration of these Omics data types presents opportunities to understand the complex sets of interactions that lead to COPD pathology. However, for the effective application of integrative Omics analysis in COPD (and all complex diseases), several important challenges need to be addressed. As these challenges are met, exciting opportunities will follow for new pathobiological insights into COPD that may lead to improved diagnostic classification, prognostic accuracy, and therapeutic development in COPD.

DISCLOSURES

Dr. Silverman reports grants from NIH during the conduct of the study; and grants from GlaxoSmithKline and Bayer outside the submitted work.

Dr. Hobbs reports grants from NIH during the conduct of the study.

ACKNOWLEDGMENTS

This work was supported by NIH Grants U01 HL089856, P01 HL114501, R01 HL133135, R01 HL147148, and R01 HL137927 to EKS and K08 HL136928 to BDH.

REFERENCES

1. Ragland MF, Benway CJ, Lutz SM et al. Genetic Advances in Chronic Obstructive Pulmonary Disease: Insights from COPDGene. Am J Respir Crit Care Med. 2019;200:677-90.
2. Loscalzo J, Barabasi AL, Silverman EK, editors. Network Medicine: Complex Systems in Human Disease and Therapeutics. Cambridge, MA: Harvard University Press; 2017.
3. Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr Opin Biotechnol. 2019;58:129-36.
4. Lahnemann D, Koster J, Szczurek E et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31.
5. Joo J, Hobbs BD, Cho MH, Himes BE. Trait Insights Gained by Comparing Genome-Wide Association Study Results using Different Chronic Obstructive Pulmonary Disease Definitions. AMIA Jt Summits Transl Sci Proc. 2020;2020:278-87.
6. Robinson PN. Deep phenotyping for precision medicine. Hum Mutat. 2012; 33:777-80.
7. Yehia L, Eng C. Largescale population genomics versus deep phenotyping: Brute force or elegant pragmatism towards precision medicine. NPJ Genom Med. 2019;4:6.
8. Silverman EK. Genetics of COPD. Annu Rev Physiol. 2020;82:413-31
9. Silverman EK, Sandhaus RA. Clinical practice. Alpha1-antitrypsin deficiency. N Engl J Med. 2009;360:2749-57.
10. Corbett E, Glaisyer H, Chan C, Madden B, Khaghani A, Yacoub M. Congenital cutis laxa with a dominant inheritance and early onset emphysema. Thorax. 1994;49:836-7.
11. Sakornsakolpat P, Prokopenko D, Lamontagne M et al. Expanded genetic landscape of chronic obstructive pulmonary disease reveals heterogeneous cell type and phenotype associations. Nature Genetics. 2019;51:494-505.
12. Silverman EK. Applying Functional Genomics to Chronic Obstructive Pulmonary Disease. Ann Am Thorac Soc. 2018;15(Suppl 4):S239-S42.
13. Baranski TJ, Kraja AT, Fink JL et al. A high throughput, functional screen of human Body Mass Index GWAS loci using tissue-specific RNAi Drosophila melanogaster crosses. PLoS Genet. 2018;14:e1007222.
14. Moll M, Sakornsakolpat P, Shrine N, et al. Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. Lancet Respir Med 2020;8:696-708.
15. Zhou X, Baron RM, Hardin M et al. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. Hum Mol Genet. 2012;21:1325-35.
16. Castaldi PJ, Guo F, Qiao D et al. Identification of Functional Variants in the FAM13A Chronic Obstructive Pulmonary Disease Genome-Wide Association Study Locus by Massively Parallel Reporter Assays. Am J Respir Crit Care Med. 2019;199:52-61.
17. Parker MM, Hao Y, Guo F et al. Identification of an emphysema-associated genetic variant near TGFB2 with regulatory effects in lung fibroblasts. Elife. 2019;8.
18. Castaldi PJ, Boueiz A, Yun J et al Machine Learning Characterization of COPD Subtypes: Insights From the COPDGene Study. Chest. 2020;157:1147-57.
19. Park J, Hobbs BD, Crapo JD et al. Subtyping COPD by Using Visual and Quantitative CT Imaging Features. Chest. 2020;157:47-60.
20. Chang Y, Glass K, Liu YY, Silverman EK et al. COPD subtypes identified by network-based clustering of blood gene expression. Genomics. 2016;107:51-8.
21. Obeidat M, Nie Y, Fishbane N et al. Integrative Genomics of Emphysema-Associated Genes Reveals Potential Disease Biomarkers. Am J Respir Cell Mol Biol. 2017;57:411-8.
22. Sun W, Kechris K, Jacobson S et al. Common Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. PLoS Genet. 2016;12: e1006011.
23. Morrow JD, Glass K, Cho MH et al. Human Lung DNA Methylation Quantitative Trait Loci Colocalize with Chronic Obstructive Pulmonary Disease Genome-Wide Association Loci. Am J Respir Crit Care Med. 2018;197:1275-84.

24. Suhre K, Raffler J, Kastenmuller G. Biochemical insights from population studies with genetics and metabolomics. Arch Biochem Biophys. 2016;589:168-76.
25. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648-60.
26. Battle A, Mostafavi S, Zhu X et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24:14-24.
27. Giambartolomei C, Vukcevic D, Schadt EE et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383.
28. Jansen R, Hottenga JJ, Nivard MG et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Hum Mol Genet. 2017;26:1444-51.
29. Sun BB, Maranville JC, Peters JE et al. Genomic atlas of the human plasma proteome. Nature. 2018;558:73-9.
30. Lamontagne M, Berube JC, Obeidat M et al. Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. Hum Mol Genet. 2018;27:1819-29.
31. Sakornsakolpat P, Morrow JD, Castaldi PJ et al. Integrative genomics identifies new genes associated with severe COPD and emphysema. Respir Res. 2018;19:46.
32. Yonchuk JG, Silverman EK, Bowler RP et al. Circulating soluble receptor for advanced glycation end products (sRAGE) as a biomarker of emphysema and the RAGE axis in the lung. Am J Respir Crit Care Med. 2015;192:785-92.
33. Chan LY, Mugler CF, Heinrich S, Vallotton P, Weis K. Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. Elife. 2018;7:e32536.
34. Shapiro SD, Endicott SK, Province MA, Pierce JA, Campbell EJ. Marked longevity of human lung parenchymal elastic fibers deduced from prevalence of D-aspartate and nuclear weapons-related radiocarbon. J Clin Invest. 1991;87:1828-34.
35. Wang D, Liem DA, Lau E et al. Characterization of human plasma proteome dynamics using deuterium oxide. Proteomics Clin Appl. 2014;8:610-9.
36. Robinson JL, Kocabas P, Wang H et al. An atlas of human metabolism. Sci Signal. 2020;13(624): eaaz1482.
37. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. 2013;20:274-81.
38. Sonawane AR, Weiss ST, Glass K, Sharma A. Network Medicine in the Age of Biomedical Big Data. Front Genet. 2019;10:294.
39. Silverman EK, Schmidt H, Anastasiadou E et al. Molecular networks in Network Medicine: Development and applications. Wiley Interdiscip Rev Syst Biol Med. 2020:e1489.
40. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169:1177-86.
41. Eicher T, Kinnebrew G, Patt A Et al Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources. Metabolites. 2020;10(5):202.
42. Sun YV, Hu YJ. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. Adv Genet. 2016;93:147-90.
43. Arneson D, Shu L, Tsai B, Barrere-Cain R, Sun C, Yang X. Multidimensional Integrative Genomics Approaches to Dissecting Cardiovascular Disease. Front Cardiovasc Med. 2017;4:8.
44. Hawe JS, Theis FJ, Heinig M. Inferring Interaction Networks From Multi-Omics Data. Front Genet. 2019;10:535.
45. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83.
46. Wang B, Mezlini AM, Demir F et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333-7.
47. Liu H, Zhao R, Fang H, Cheng F, Fu Y, Liu YY. Entropy-based consensus clustering for patient stratification. Bioinformatics. 2017;33:2691-8.
48. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
49. Faner R, Morrow JD, Casas-Recasens S et al. Do sputum or circulating blood samples reflect the pulmonary transcriptomic differences of COPD patients? A multi-tissue transcriptomic network META-analysis. Respir Res. 2019;20:5.
50. Shi WJ, Zhuang Y, Russell PH et al Unsupervised discovery of phenotype-specific multi-omics networks. Bioinformatics. 2019;35:4336-43.
51. Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing messages between biological networks to refine predicted interactions. PLoS ONE. 2013;8:e64832
52. Moll M, Qiao D, Regan EA et al. Machine Learning and Prediction of All-Cause Mortality in Chronic Obstructive Pulmonary Disease. Chest. 2020; 158:952-64.
53. Keller A, Fehlmann T, Ludwig N et al. Genome-wide MicroRNA Expression Profiles in COPD: Early Predictors for Cancer Development. Genomics Proteomics Bioinformatics. 2018;16:162-71.
54. Kim S, Herazo-Maya JD, Kang DD et al. Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. BMC Genomics. 2015;16:924.
55. Kusko RL, Brothers JF, 2nd, Tedrow J et al. Integrated Genomics Reveals Convergent Transcriptomic Networks Underlying Chronic Obstructive Pulmonary Disease and Idiopathic Pulmonary Fibrosis. Am J Respir Crit Care Med. 2016;194:948-60.
56. Huang GT, Athanassiou C, Benos PV. mirConnX: condition-specific mRNA-microRNA network integrator. Nucleic Acids Res. 2011;39(Web Server issue): W416-23.
57. Li CX, Wheelock CE, Skold CM, Wheelock AM. Integration of multi-omics datasets enables molecular classification of COPD. Eur Respir J. 2018;51.
58. Mastej E, Gillenwater L, Zhuang Y, Pratte KA, Bowler RP, Kechris K. Identifying Protein-metabolite Networks Associated with COPD Phenotypes. Metabolites. 2020;10.
59. Zhou W, Nielsen JB, Fritsche LG et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50:1335-41.
60. Zhou W, Sailani MR, Contrepois K et al. Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature. 2019;569:663-71.