iScience

CellPress
OPEN ACCESS

Perspective

A panoramic view of proteomics and multiomics in precision health

Mara Zilocchi,[1] Cheng Wang,[2] Mohan Babu,[1,*] and Jingjing Li[2,*]

SUMMARY

Health is often qualitatively defined as a status free from disease and its quantitative definition requires finding the boundary separating health from pathological conditions. Since many complex diseases have a strong genetic component, substantial efforts have been made to sequence large-scale personal genomes; however, we are not yet able to effectively quantify health status from personal genomes. Since mutational impacts are ultimately manifested at the protein level, we envision that introducing a panoramic proteomic view of complex diseases will allow us to mechanistically understand the molecular etiologies of human diseases. In this *perspective* article, we will highlight key proteomic approaches to identify pathogenic mutations and map their convergent pathways underlying disease pathogenesis and the integration of omics data at multiple levels to define the borderline between health and disease.

INTRODUCTION

As early as in 1996, the practice of genomic medicine had been precisely described by Laurie Garrett, a Pulitzer prize-winning writer, where she imagined that by the year 2020, everyone will have his/her personal genome information in a wallet-sized card guiding physicians' clinical decision (Garrett, 1996). Compared with the half-completed Human Genome Project in 1996, today we have completed whole-genome sequencing for millions of individuals and have identified numerous novel disease-associated loci across the genome. Technological advancements are now profoundly transforming our clinical practice, where tumor sequencing and non-invasive prenatal testing have become routine clinical procedures. However, to the general population, the imaginary scenario made in 1996 has not yet come true: while we are now able to make a genome card for each person, to a large extent, we are still unable to effectively translate the massive genomic information into clinical knowledge.

The widely used approach to analyze complex disease genomes is genome-wide association studies (GWAS), which has identified thousands of loci associated with numerous diseases (NHGRI-EBI GWAS Catalog). However, for many diseases, even with a very strong genetic component, GWAS often yield no signal (Manolio et al., 2009), leading to a decade-long search for the "missing heritability''. Built on GWAS, the recently developed polygenic risk score model has demonstrated predictive power for few diseases (Khera et al., 2018) but is not broadly applicable to most disease types. By scanning each individual genetic locus across the genome one at a time, GWAS identify at-risk loci in diseases that display imbalanced allelic frequencies between cases and controls, thereby tagging haplotypes by sentinel single nucleotide polymorphisms (SNPs). The disparities in allelic frequencies do not directly indicate functional consequences of genomic variants; as such it is challenging to identify causal variants from regular GWAS investigations. If we are able to quantify the molecular effect for every base change across the genome, integrating functional data into the existing GWAS framework would substantially boost the power of detecting at-risk loci and naturally unveil causal variants in diseases, thereby fundamentally innovating our interpretation of genomic variants in human pathologies.

Another challenge with GWAS is mutational heterogeneity in complex human diseases, where hundreds or thousands of genes are implicated in a given disease and different patients usually carry different clinical mutations (Figures 1A and 1B). Because GWAS assumes independent contribution from every single locus to disease etiologies, the intrinsic heterogeneity would become a bottleneck for identifying individual disease-associated variants that are more common and prevalent in cases relative to controls. To capture these variants, large sample sizes are naturally needed to enrich disease-associated alleles in patient

[1]Department of Biochemistry, University of Regina, Regina, Saskatchewan S4S 0A2, Canada

[2]The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, the Bakar Computational Health Sciences Institute, the Parker Institute for Cancer Immunotherapy, and the Department of Neurology, School of Medicine, University of California, San Francisco, CA, USA

*Correspondence: mohan.babu@uregina.ca (M.B.), jingjing.li@ucsf.edu (J.L.)

https://doi.org/10.1016/j.isci.2021.102925

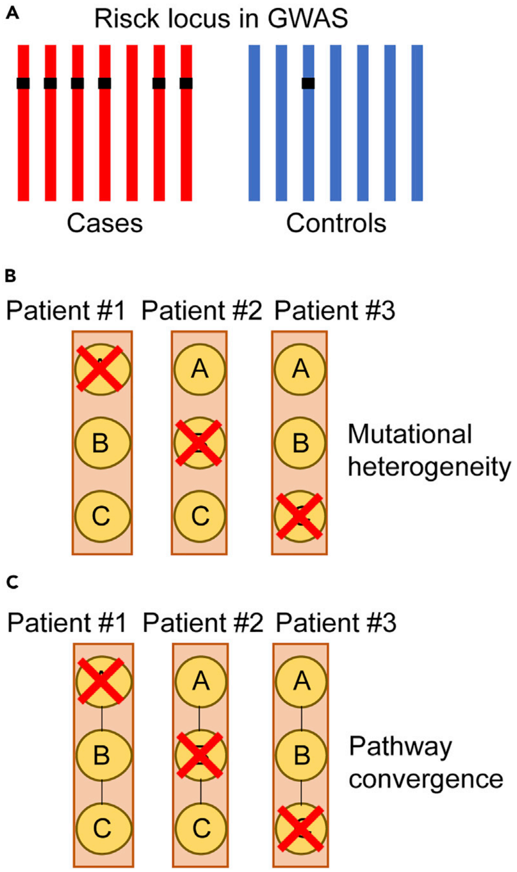

Figure 1. Mutational heterogeneity and pathway convergence

(A) GWAS examine each locus for the allelic enrichment in cases relative to controls.

(B) GWAS are often challenged by mutational heterogeneity. For example, patients #1, #2, and #3 have different affected genes (A, B and C) leading to the same disease.

(C) If the affected genes A, B, and C are on the same pathway, the mutational heterogeneity should be understood as mutational convergence.

populations to reach genome-wide significance (Figure 1B). However, this model might not best reflect the biology of human diseases, which are hallmarked by perturbations on molecular pathways (Nesterova et al., 2019; Fruman et al., 2017) (Figure 1C). For example, prostate cancer is often characterized by perturbation on androgen signaling pathways, and obstetric conditions are often linked with defective progesterone signaling. In other words, individual patients might carry highly heterogeneous mutations, but these mutations in fact converge onto common biological pathways. As such, diseases with overlapping molecular components tend to develop similar symptoms (Menche et al., 2015; Zhou et al., 2014). Therefore, going beyond individual mutations and bringing systems thinking will help innovate our genome analysis by identifying mutationally convergent pathways, which would effectively circumvent the challenge from extreme mutational heterogeneity in human diseases.

When it comes to functional determination of individual mutation consequences or identifying mutational convergence onto biological pathways, proteomic profiling is apparently the most straightforward solution. In this article, we will present our perspectives on the next generation of disease genome analysis by introducing large-scale proteomic techniques and systems biology approaches. At the individual mutation level, we will demonstrate the use of proteomic analysis to empower our genome analysis by quantifying mutational effects on altering protein abundance and function. At the systems level, we will particularly demonstrate the power of proteomic profiling in charting the cellular maps underlying human diseases, which will enable us to identify disease-associated pathways and will provide us with a network view of complex diseases. Lastly, we will illustrate the use of machine learning approaches to integrate proteomic profiling and large-scale genome analysis, enabling a quantitative computation of the borderline between human health and disease. In particular, the integration of different omics techniques such as single-cell RNA sequencing with the single-cell proteomics will have a deep impact on the interpretation of different disease phenotypes by profiling the transcript and protein expression of single cells, elucidating the heterogeneity of human biopsy samples, and highlighting susceptibilities in specific cell types. Single-cell metabolomic represents another important asset for the prediction or classification of human diseases. The characterization of small molecules is necessary to uncover the metabolic changes inside different cell types and to understand how the alteration of specific metabolic pathways alters the entire cellular metabolism and genetic expression (Kiviet et al., 2014). The application and the integration of these single-cell technologies will therefore allow a more accurate characterization of the borders between health and disease conditions.

DETERMINING MUTATIONAL CONSEQUENCES IN ALTERING PROTEIN ABUNDANCE

Recent studies have consistently revealed that more than 90% of at-risk loci in complex human diseases fall in non-coding regions (Schaub et al., 2012; Corradin and Scacheri, 2014; Boix et al., 2021). As such, disrupting regulatory elements so as to perturb gene regulation represents a major mechanism underlying human diseases. The Genotype-Tissue Expression (GTEx) project systematically cataloged human genomic variation associated with gene expression across multiple tissue types (e.g. expression quantitative loci, eQTLs), which have provided an entry point connecting genetic changes with disease predispositions utilizing mRNA gene expression as an endophenotype. However, the end product of gene regulation is protein, and cellular protein abundance could be highly disproportionate from mRNA expression given post-transcriptional regulation mechanisms and environmental contribution (Jiang et al., 2020; Liu et al., 2016). This notion was particularly highlighted by the landmark paper of the quantitative proteome map of the human body (Jiang et al., 2020), where numerous genes with ubiquitous mRNA expression across tissues in fact displayed strong tissue specificities in their protein abundance. Therefore, ideally the consequences of human genomic mutations should be understood at the protein level, and the recent rapid development of proteomic technologies has made this possible.

Given the number of biological factors involved in the regulation of protein expression and the importance of protein abundance in describing the physiological or pathological state of a biological system, several proteomic techniques have been developed to quantitatively evaluate the proteome. In this context, two different strategies can be adopted to quantify the protein abundance in human samples: (1) stable isotope labeling techniques; and (2) label-free methods. Despite each of these techniques having its own pros and cons in the evaluation of the human proteome (DeSouza and Siu, 2013; Aly et al., 2021), the rapidly evolving field of precision medicine has driven the development of the approach known as protein Quantitative Trait Loci (pQTLs) (Ye et al., 2020), which is able to correlate the genetic variant with the protein abundance and the relative clinical trait or disease risk, thus elucidating the causal role of that protein in a specific disease state. Wu and colleagues, for example, used isobaric tandem mass tag-based quantitative mass spectrometry to determine protein levels of ~6,000 genes in lymphoblastoid cell lines (LCLs) from 95 ethnically diverse individuals genotyped in the HapMap Project, leading to a discovery of numerous *cis*-pQTLs across the genome, whose allelic alterations were significantly associated with protein abundance of their neighboring genes in LCLs (Wu et al., 2013). These genetic loci thus provided us a glimpse into the genetic control of protein abundance in humans, enabling us to directly interpret mutational consequences at the protein level. A more recent work (Robins et al., 2021) further illustrated the power of integrating proteomics in our current genomic analysis, where numerous pQTLs for 7,376 proteins were identified by performing proteomic profiling in 330 dorsolateral prefrontal cortical samples. Leveraging these brain pQTL loci, a follow-up analysis further re-assessed potential functional implications of genomic loci in large-scale GWAS of seven neurological phenotypes (Alzheimer disease, amyotrophic lateral sclerosis, depression, insomnia, intelligence, neuroticism, and schizophrenia), and identified putative causal loci underlying these conditions (Kibinge et al., 2020). In a similar effort, plasma pQTLs were also identified (Sun et al., 2018), which enabled us to directly identify at-risk loci from the genome for cardiovascular disease (Yao et al., 2018). Taken together, integrating proteomic profiling techniques has now allowed us to directly associate genetic allelic changes with alterations of protein abundance, providing a direct indicator to disease predispositions and suggesting targets for prevention and medical intervention.

Lastly, the rise of single-cell RNA sequencing technique paved the way for the optimization of single-cell proteomic approaches to accurately quantify proteins from a small number of cells (Kelly, 2020). Indeed, despite the gap that still exists between sample size and proteome coverage, recent technological advancements have allowed us to reliably profile thousands of proteins starting from just a few hundred of cells (Zhu et al., 2018; Budnik et al., 2018; Dou et al., 2018). While the field of single-cell proteomics is still not completely mature, its continuous development will allow the spatial resolution of mammalian samples, thus uncovering the protein abundance heterogeneity of complex tissues.

DETERMINING MUTATIONAL CONSEQUENCES IN ALTERING PROTEIN FUNCTION

While proteomic approaches can be leveraged to directly determine the effect of mutations on the alteration of protein abundance, recent methodological advancement has enabled us to determine mutational effects on changing protein structure and interaction. The 3D protein structures are determined by their amino acid sequence, which in turn is genetically controlled. Therefore, discovering the relationship between genome sequencing data, amino acid sequence, and protein structure is indispensable for understanding the molecular basis of complex human diseases. Over the years, many efforts have been undertaken to resolve the 3D structure of proteins via X-ray crystallography, high-resolution nuclear magnetic resonance spectroscopy, cryogenic electron microscopy, and protein-modeling algorithms (Kuhlman and Bradley, 2019). These techniques assess non-synonymous mutations (loss-of-function and missense mutations) and are expected to pave the way for advancing our understanding of mutations from exome-sequencing data for complex human diseases. Compared with no nonsense mutations resulting in premature stop codons, characterizing functional consequences of missense mutations (amino acid replacement) are often challenging.

Classical structural biology methods are employed to determine alterations in physicochemical properties (e.g. difference in solvent exposure between wild-type and mutated alleles) associated with amino acid replacement, which are integrated with evolutionary conservation analysis to derive a quantitative metric to score mutational deleteriousness from exome-sequencing data. Interestingly, as opposed to our initial expectation of mutational effects on affecting structure or stability of individual proteins, increasing evidence has now shown that disease-causing missense mutations are likely to disrupt protein-protein interactions. A recent work (Fragoza et al., 2019) leveraged the yeast-two-hybrid (Y2H) platform to experimentally assess mutational consequences of missense mutations from whole-exome-sequencing data on perturbing protein interactions, and estimated that ~10.5% among all missense variants per genome likely affect protein interactions, instead of causing unstable protein expression. This observation is important because the perturbation of protein interactions simply implies substantial rewiring of the cellular protein-protein interaction network, which likely contribute to the diversified phenotypes, especially disease predispositions, among the human population. More importantly, this study further revealed that those missense mutations disrupting protein interactions are widely distributed across the entire allele frequency spectrum; as such, perturbing protein interactions likely represent common mechanisms underlying human diseases. Another recent work further strengthened these concepts by directly mapping genomic mutations onto protein sequences encoding residues in protein-protein interaction interfaces, and indeed observed that both germline mutations in human diseases and somatic mutations in different cancer types were more likely to affect residues in protein-protein interaction interfaces (Cheng et al., 2021). Focusing on cancer, intriguingly, these affected protein interactions were highly correlated with clinical outcomes (survival) and drug responses.

Recently, a study conducted (Wierbowski et al., 2020) on the protein-protein interactions between SARS-CoV-2 and its human host and the impact of human genetic variants on the disruption of native protein-protein interactions and, consequently, on patient symptoms and responses following SARS-CoV-2 exposure, was able to generate a web interface containing a 3D structural interactome meta-analysis and the prediction of the binding between viral-human interaction interfaces and drug repurposing candidates. Given all these findings, proteomic techniques (protein abundance profiling, Y2H and pull-down assay) to define protein interactomes and 3D protein structural modeling to spatially resolve localization of genomic mutations, have apparently opened a new avenue to future genomic medicine, from disease prognostics to treatment strategy development.

MAPPING THE PROTEIN INTERACTOME TO REVEAL MUTATIONALLY CONVERGENT PATHWAYS

Complex diseases are hallmarked by locus heterogeneity, where different patients carry different sets of mutations. However, it is important to emphasize that the concept of locus heterogeneity was simply derived from the one-dimensional genome, and viewing these mutations from the proteomic dimension gives a different view. In this context, affinity purification assays and Y2H screenings offer a unique opportunity to uncover the protein interactions impact on complex human diseases. While Y2H assays can only verify the physical interaction in protein pairs, the affinity purification methods are able to identify direct or indirect interactions among a group of proteins, thus representing one of the most common methods used to characterize complex protein interactions in different model systems. Over the years, pull-down assays have also been applied to uncover the impact of extreme locus heterogeneity on protein associations. For example, our recent work (Li et al., 2015) on co-immunoprecipitation coupled with mass spectrometry in neuronal cells revealed proteins interacting with key autism spectrum disorder (ASD) proteins that were individually identified from clinical studies. Intriguingly, despite these ASD proteins found from disparate patients with idiopathic and syndromic ASD, their interacting proteins formed a highly connected network, suggesting a mutational convergence onto molecular pathways in complex diseases. It is also the same for the well-known BAF complexes and PI3K/Akt/mTOR pathway, where almost all their protein members had been individually identified as at-risk loci in neurodevelopmental disorders. Moreover, the recent resolution of the mitochondrial protein complex reorganization during neuronal differentiation unveiled the uncharacterized protein C20orf24, whose 3′ UTR variant is associated with mitochondria respiratory chain complex deficiency in patients, as a respirasome assembly factor, as well as showing that the binding between NENF and the Parkinson's disease-associated proteins DJ1 and PINK1 resulted in the improvement of neurotrophic activity and, consequently, neuronal survival (Moutaoufik et al., 2019). Additionally, we showed that the SOD1-PRDX5 interaction, critical for mitochondrial redox homeostasis, can be perturbed by amyotrophic lateral sclerosis-linked SOD1 allelic variants, and also established a functional role for neurodegenerative-linked factors coupled with IkBε in nuclear factor-kB (NF-kB) activation (Malty et al., 2017). In addition to neurological disorders, our systems biology analyses revealed mutational convergence in cardiovascular or pulmonary diseases and prostate cancer (Li et al., 2018; Wang and Li, 2020). Deciphering complex human diseases through the integration of genomics, transcriptomics, and proteomics data has also made challenging by the need of using cell lines or mouse models to investigate the mechanisms contributing to a specific pathology. Indeed, obtaining genetic information from a patient simply require a blood or a saliva sample, while dissecting the transcriptomic or the proteomic impact of a mutation in a specific tissue often needs the use of models that are not able to recapitulate the complex human disease phenotypes. This important gap in translating the discoveries made in well characterized cellular or animal models to heterogeneous human diseases drove the development of more close-to-natural samples, like induced pluripotent stem cell (iPSC)-derived cell lines and organoids (Zilocchi et al., 2020).

Overall, including a proteomic perspective to the genomic information will help revealing the convergent disease pathways from heterogeneous mutations. As such, genetic architecture of human phenotypes is best characterized by mapping genomic mutations onto molecular pathways, calling for large-scale proteomic profiling to derive a complete cellular map of protein-protein interaction. Toward this goal, investigators have formed the Psychiatric Cell Map Initiative (Willsey et al., 2018) to systematically delineate a protein-protein interaction map underlying brain development, which will not only identify molecular pathways where pathological mutations converge, but will also reveal the shared molecular etiologies at a pathway level among distinct but related neurodevelopmental disorders, such as autism, intellectual disability, epilepsy and schizophrenia. Together, integrating proteomic profiling in disease genome analysis will provide deep mechanistic insights into disease etiologies, which are absent from genomic analysis alone.

LEVERAGING DEEP LEARNING TO INTEGRATE PROTEOMIC PROFILING AND PATIENT GENOMES

Once we have a reference cellular map of protein interactions, we will be able to convert the traditional mutation analysis from the one-dimensional genome space into a machine learning problem in a high-dimensional space. We previously illustrated the problem formulation by mapping genomic mutations onto a cellular network, followed by developing graph search algorithms to identify compact sub-networks with increased mutational load in cases relative to controls ((Li et al., 2019), Figure 2). However, calculating mutational load is often challenged by unclear molecular consequences of individual genomic variants. Borrowing strength from proteomic profiling data, and the accumulating pQTL data that have become available in different human tissue and cell types will now allow the estimation of tissue-specific mutational pathogenicity. These functional data (i.e., quantifying mutational effects on protein abundance, structure and interaction as described above) will enable us to directly aggregate the impact of pathogenic

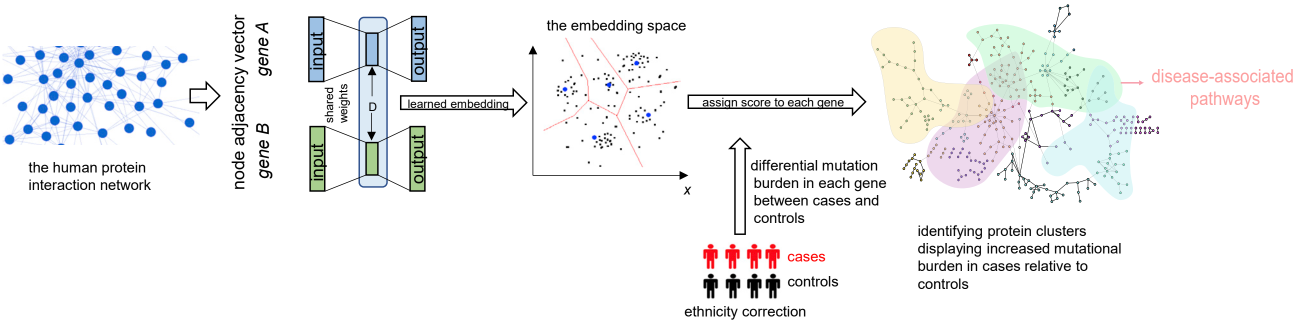
the human protein interaction network
node adjacency vector
gene A
gene B
input
output
input
output
shared weights
D
learned embedding
the embedding space
x
assign score to each gene
differential mutation burden in each gene between cases and controls
cases
controls
ethnicity correction
identifying protein clusters displaying increased mutational burden in cases relative to controls
disease-associated pathways

Figure 2. Structural deep network embedding (SDNE) to identify disease-associated pathways from large-scale disease genomes
The embedding technique is to map proteins on the network onto a low-dimensional space where the local topological features of a given node protein are still preserved. SDNE uses two deep AutoEncoder networks, takes the adjacency vector of each protein (network node) as input, and learns its embeddings in a new space with compressed dimension. For a given protein in the embedding space, SDNE preserves the first and second order topological neighbors from the original network. Clustering proteins in the embedding space will identify functionally related proteins, thereby defining biological pathways. For each gene (protein), we then define its mutation burden by aggregating common, rare and de novo deleterious variants in the gene (weighted aggregation), followed by a comparison with the expected mutational burden in control samples. The test statistic is then assigned to each protein to reflect the differential mutational burden of a given gene in cases relative to controls. Lastly, we will identify protein clusters displaying significantly increased mutational burden in cases relative to controls. These clusters will be mapped back onto the original interaction network for their physical organization to represent biological pathways.

mutations on each gene at the protein level. We can then score each protein based on their differential impact scores in cases relative controls, followed by identifying the underlying pathways on which the seemingly heterogeneous mutations converge.

We herein introduce a deep learning theoretical framework to uncover genetic basis from large-scale disease genomes by integrating proteomic information. Compared with our earlier proposal based on graph search algorithms (Li et al., 2019), which are often computationally challenging when handling large-scale biological networks and millions of genomic mutations, the deep learning framework is computationally scalable and efficient. We introduce the Structural Deep Network Embedding (SDNE) framework (Wang et al., 2016) for disease genome analysis. SDNE maps each protein from a large biological network onto a compressed space (embedding) by implementing semi-supervised deep autoencoder networks (Figure 2). Therefore, unlike PCA, the SDNE mapping is non-linear and guarantees that the geometric distribution of the proteins in the new space preserves their local topological structure on the biological network. Clustering proteins in the new embedding space (Figure 2) will reveal hidden pathway structures not seen in the original topological network (Wang et al., 2016). SDNE well solves the computing scalability problem, and the incomplete information on biological networks will be addressed in the new "embedding" space, where proteins in related pathways will be geometrically clustered together despite their missing links on the original network. After clustering, we can map mutations onto each protein in the embedding space: each protein will be assigned with a score quantifying differential burden of aggregated consequential mutations between cases and controls. Common or rare variants could be analyzed separately or based on their weighted aggregation (Curtis, 2019). With this score assignment, the clusters enriched for proteins frequently affected by pathogenic mutations in cases relative to controls can be easily identified by regular statistical tests. Mapping the identified protein clusters back onto the original interaction network will identify their physical organization, which naturally reveals disease-associated pathways (Figure 2).

Although the theoretic model we presented above is an unsupervised model solely driven by clustering structure of pathogenic mutations, the model can be easily extended to supervised models guided by a list of known disease-associated genes as training data. Krishnan and colleagues previously developed a classifier trained on genes known to be implicated in autism, which was utilized to score all human candidate genes in autism. The classifier was constructed based on topological similarity between any given genes and known autism genes on a defined brain-specific gene-gene association network (Krishnan et al., 2016). Although their original model was not designed for analyzing personal genomes, the recent development of a graph convolutional network has made it possible, where each protein on the interaction network could be represented by a feature vector encoding its topological position, enrichment of pathogenic mutations, and gene expression. The deep learning algorithm will be trained to identify proteins in close proximity to known disease genes in the feature space.

FROM DEEP MOLECULAR PROFILING TO DEEP PHENOTYPING: THE EMERGING TECHNOLOGICAL DEVELOPMENT

While leveraging deep learning to integrate multi-omics data could help address many long-standing challenges in our genome analysis, one fundamental problem that has remained is the extreme phenotypic heterogeneity. In a typical clinical analysis, we classify human populations into cases and controls; however, the borderline between health and disease is often not binary but personal and dynamic. For example, the normal body temperature (oral) varies from 33.2°C to 38.2°C among the human population (Sund-Levander et al., 2002), and therefore adopting a single threshold based on the population average would likely result in erroneous decision making in our clinical practice. This personalized perception of human diseases and treatments has guided the shifting from the "one-size-fits-all" approach of the epidemiological studies, to the more accurate precision medicine field, which aims at defining a custom-made diagnosis and treatment for each patient. In particular, by pairing large multidimensional biological datasets, which capture individual variabilities associated with a peculiar phenotype, and artificial intelligence algorithms, precision medicine allows the prediction of disease risk, treatment response, and other outcomes in each subject based on their own physio-pathological characteristics (Uddin et al., 2019).

Electronic health records (EHRs) contain rich information about patient journey during their clinical visits, and deep learning models have been recently developed (Morel et al., 2020; Jaotombo et al., 2020; Desautels et al., 2017) to predict future clinical events (e.g., in-hospital mortality and unplanned readmission) from past events recorded in individual's EHRs. The UK-Biobank, TOPMed among many other biobank and consortium resources have now matched each patient's genome (exome) with EHRs, enabling genome scan for multiple traits at the same time. Our recent work (Li et al., 2018) leveraged a machine learning framework to integrate patients' personal genomes, EHR data, and personal lifestyles, which by aggregation accurately predicted disease occurrences of abdominal aortic aneurysm, a cardiovascular condition prevalent among aged population. It is important to note that genomic mutation profiles predispose individuals to a given disease, whereas EHR and lifestyle data predict disease risk in near term. By aggregating these two elements, it is possible to accurately predict the disease risk as we demonstrated before (Li et al., 2018). Going beyond disease risk prediction, one can further seek clinically actionable solutions to reduce disease risk by modifying lifestyles conditioned on one's personal genome.

Despite the clinical utility of EHR data for stratifying patient phenotypes, the data itself are sparse and are only recorded during discrete clinical visits. We and several other groups are actively exploring alternative solutions to achieve deep and precise phenotyping (Li et al., 2019). Wearable technologies represent a proactive solution, which by integrating with cloud computing and storage can easily achieve physiological data acquisition in real time. Clinical diagnostic decisions could be immediately made by analyzing the acquired physiological data (heart rate, skin temperature, respiratory rate, etc.). For example, recent work repurposed consumer smartwatches to longitudinally track physiological signal fluctuations of study participants and successfully achieved symptomatic and pre-symptomatic detection of COVID-19 (Quer et al., 2021; Mishra et al., 2020). Similarly, investigators have also used continuous glucose monitoring for dense sampling to longitudinally track glucose dynamics and uncovered highly personal glucotypes among study participants (Hall et al., 2018), thereby suggesting the limitation of using single-time-point measurements in the existing clinical diagnosis practice. More importantly, the observation also calls for fine stratification of diabetic patients based on personal glucose profiles to identify molecular mechanisms specific to personal glucotypes. By enabling personal deep phenotyping, we envision that multi-omics data profiling and integration will be ultimately achieved at a personal level.

LOOKING FORWARD: FINDING THE BORDERLINE BETWEEN DISEASE AND HEALTH LEVERAGING BIGGER DATA, BETTER MODEL, AND NEWER TECHNOLOGIES

We live in a changing world, where since the past decade, we have witnessed the rapid: (1) growth of high-throughput technologies, (2) replacement of microarrays by RNA sequencing, (3) accumulation of disease genomes, and (4) development of single-cell and proteomic technologies, enabling us to investigate cellular events at an unprecedented resolution. Compared with technological advancements, our

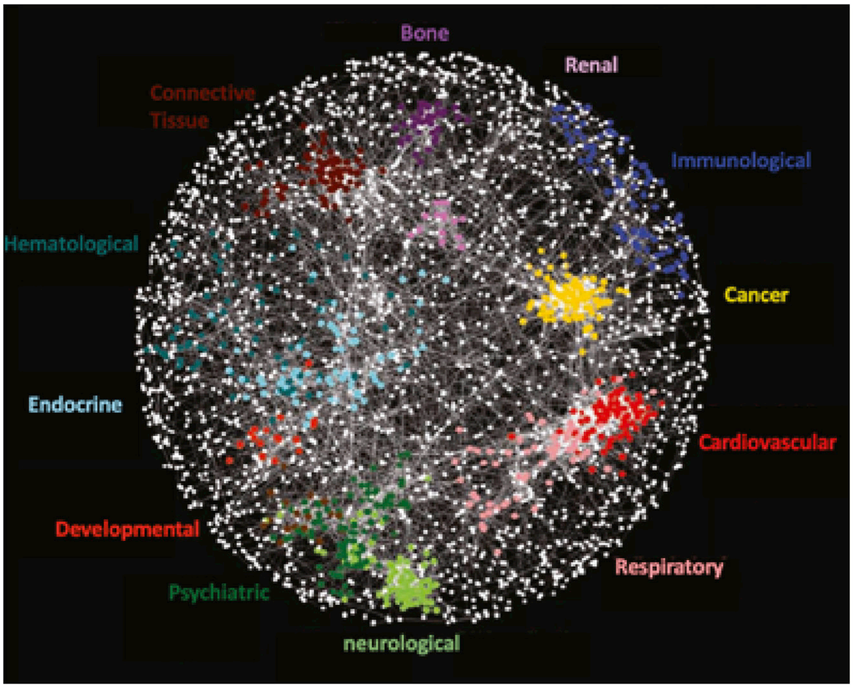

Figure 3. A conceptual biological network connecting proteins with physical or genetic interactions
Proteins associated with related pathways are proximal to each other, while proteins with non-overlapped biological functions are positioned far apart. Colored regions indicate enriched disease-associated pathways for different categories. Figure is from Science. 2010; 327:425-431. Reprinted with permission from AAAS.

analytical frameworks have remained largely unchanged, where we continue to perform GWAS and the search for missing heritability has remained elusive. We reasoned that these statistical genetic models infer disease associations purely based on allele frequencies, which do not directly model molecular functions. Therefore, genetic heterogeneity has remained a major challenge (Figure 1B), which would require large sample size to enrich at-risk alleles so as to reach genome-wide significance.

We present our view for the nature of complex diseases on the global human protein interaction network, where genes implicated in a particular disease type are clustered in related biological pathways. Overlapping pathways between diseases likely reflect overlapping phenotypic spectra in each disease (Figure 3). Therefore, proteomic profiling is strongly desired to help elucidate mutational consequences and to construct tissue/cell-type specific cellular maps, which will substantially expediate our disease genome analysis. This integrative view of human diseases has been extensively applied by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) for the characterization of different cancer types. In particular, the employment of proteomics-based approaches allowed the elucidation of genomics alteration effects on the disease proteomics landscape and the identification of specific therapeutic targets (Krug et al., 2020; Mertins et al., 2016; Zhang et al., 2014; Wu et al., 2019). The big amount of data derived from these multi-omics studies were then incorporated in the open-source platform cBioPortal to visualize the multidimensional aspects of tumors in the context of proteomics, genomics, and clinical data (Wu et al., 2019). This integrative view requires innovative analytical models, which represents the next-generation framework for large-scale disease genome analysis. Integrating with proteomic data, we essentially map genomic mutations from the one-dimension genome to a multidimensional space, which can be easily formulated into a space embedding problem in machine learning (Figure 2). Therefore, extending the existing statistical association framework, we envision that deep learning will play a fundamental role in quantifying allelic effects and especially in network embedding to reveal disease-associated pathways. This integrative framework naturally circumvents the genetic heterogeneity challenge and converts it into finding hidden pathway structures on biological networks, a typical problem in machine learning. We envision that integrating with multi-omics data will substantially lessen the demand for large-scale patient samples but will require the development of advanced deep learning platforms, which represent an exciting new opportunity in this field.

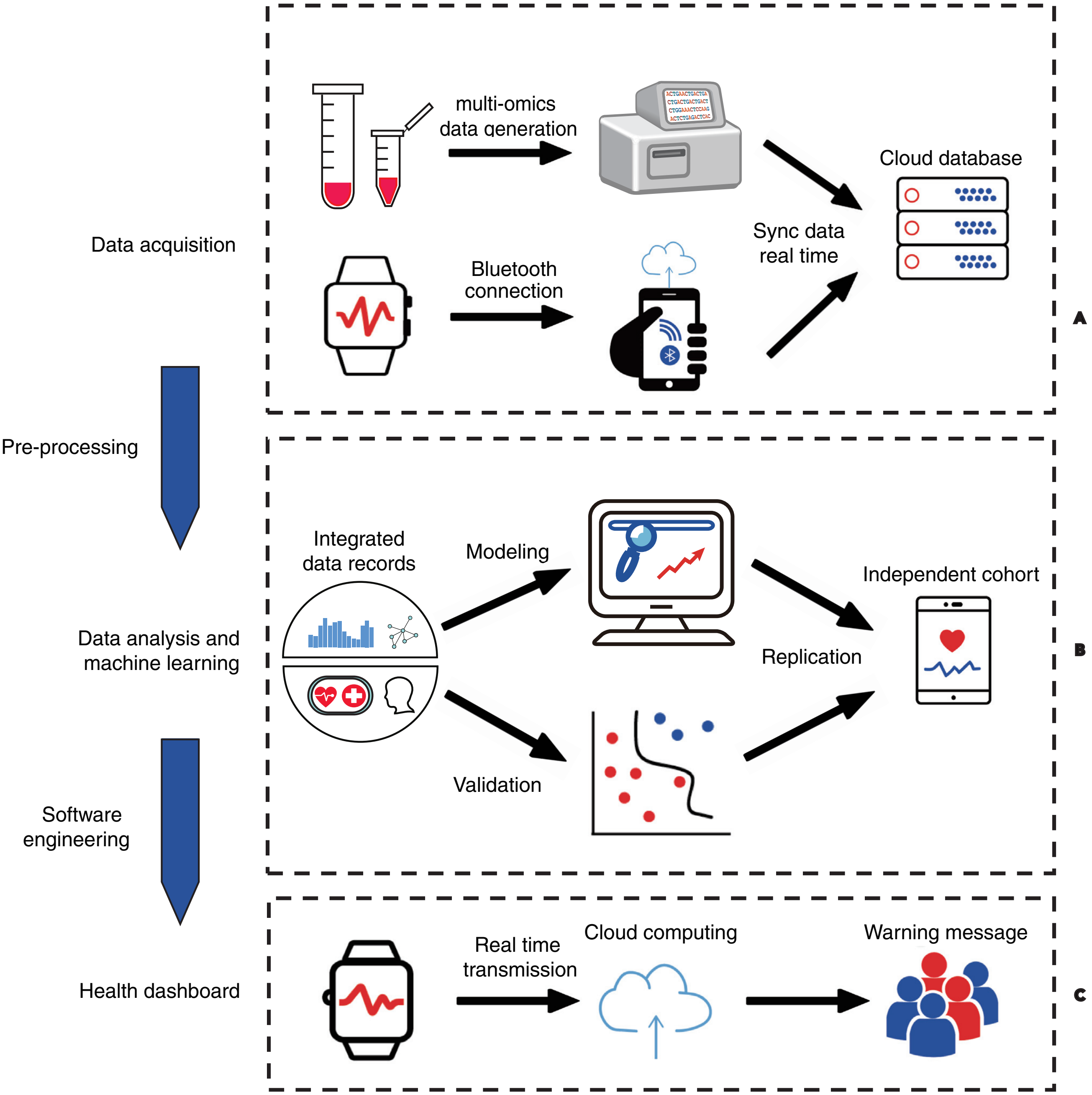

Figure 4. Working schemes of precision mapping of personal genome

(A and B) Acquisition and storage of integrated multi-omics and clinical data (A); Training, validation and replication of best machine learning models customized for each individual (B).

(C) Personalized real-time heath monitoring with transmitted data.

To achieve precision mapping from genomes onto phenotypes, we envision that mapping personal genomes onto personal traits, instead of associating population allele frequencies with population average of phenotypic traits, will revolutionize our preventative and therapeutic strategies in our clinical practice and will significantly advance our personal healthcare. While personal omics-profiling has become available (Chen et al., 2012), unified protocols and standards are required for deep and longitudinal phenotyping in future practice. Emerging technologies and 3D printing techniques have made it possible to design and manufacture many sensor types to capture physiological signals that cannot be recorded in previous research, such as wearable headband sensors longitudinally recording brain electroencephalogram signals and sensor chips woven into a shirt for proactively tracking muscle contraction events or monitoring electrocardiogram signals (Medgadget, 2020). These technologies enable longitudinal phenotyping and will help define disease subtypes and achieve stratification of patients, facilitating fine mapping of molecular components for quantitatively defined phenotypes (Figure 4).

ACKNOWLEDGMENTS

This work was supported by grants from the Canadian Institutes of Health Research (CIHR) (MOP-FDN-154318) and National Institutes of Health (R01GM106019) to M.B. J.L. acknowledges funding support from Chan Zuckerberg Initiative (the Silicon Valley Community Foundation, DAF2020-217686) and Simons Foundation (SFARI 697827, J.L.).

AUTHOR CONTRIBUTIONS

J.L. and M.B. conceived the project. All authors contributed to writing and revising the manuscript.

DECLARATION OF INTERESTS

J.L. is a cofounder and is on the advisory board of SensOmics, Inc.

REFERENCES

Aly, K.A., Moutaoufik, M.T., Phanse, S., Zhang, Q., and Babu, M. (2021). From fuzziness to precision medicine: on the rapidly evolving proteomics with implications in mitochondrial connectivity to rare human disease. iScience *24*, 102030.

Boix, C.A., James, B.T., Park, Y.P., Meuleman, W., and Kellis, M. (2021). Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature *590*, 300–307.

Budnik, B., Levy, E., Harmange, G., and Slavov, N. (2018). SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. Genome Biol. *19*, 161.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y., Chen, R., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293–1307.

Cheng, F., Zhao, J., Wang, Y., Lu, W., Liu, Z., Zhou, Y., Martin, W.R., Wang, R., Huang, J., Hao, T., et al. (2021). Comprehensive characterization of protein-protein interactions perturbed by disease mutations. Nat. Genet. *53*, 342–353.

Corradin, O., and Scacheri, P.C. (2014). Enhancer variants: evaluating functions in common disease. Genome Med. *6*, 85.

Curtis, D. (2019). A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score. Eur. J. Hum. Genet. *27*, 114–124.

Desautels, T., Das, R., Calvert, J., Trivedi, M., Summers, C., Wales, D.J., and Ercole, A. (2017). Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open *7*, e017199.

DeSouza, L.V., and Siu, K.W. (2013). Mass spectrometry-based quantification. Clin. Biochem. *46*, 421–431.

Dou, M., Zhu, Y., Liyu, A., Liang, Y., Chen, J., Piehowski, P.D., Xu, K., Zhao, R., Moore, R.J., Atkinson, M.A., et al. (2018). Nanowell-mediated two-dimensional liquid chromatography enables deep proteome profiling of <1000 mammalian cells. Chem. Sci. *9*, 6944–6951.

Fragoza, R., Das, J., Wierbowski, S.D., Liang, J., Tran, T.N., Liang, S., Beltran, J.F., Rivera-Erick, C.A., Ye, K., Wang, T.Y., et al. (2019). Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat. Commun. *10*, 4141.

Fruman, D.A., Chiu, H., Hopkins, B.D., Bagrodia, S., Cantley, L.C., and Abraham, R.T. (2017). The PI3K pathway in human disease. Cell *170*, 605–635.

Garrett, L. (1996). The dots are almost Connected....Then what? : mapping the human genetic code [online]. Los Angeles time. https://www.latimes.com/archives/la-xpm-1996-03-03-tm-42636-story.html.

Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T., and Snyder, M. (2018). Glucotypes reveal new patterns of glucose dysregulation. PLoS Biol. *16*, e2005143.

Jaotombo, F., Pauly, V., Auquier, P., Orleans, V., Boucekine, M., Fond, G., Ghattas, B., and Boyer, L. (2020). Machine-learning prediction of unplanned 30-day rehospitalization using the French hospital medico-administrative database. Medicine (Baltimore) *99*, e22361.

Jiang, L., Wang, M., Lin, S., Jian, R., Li, X., Chan, J., Dong, G., Fang, H., Robinson, A.E., Consortium, G.T., and Snyder, M.P. (2020). A quantitative proteome map of the human body. Cell *183*, 269–283.e19.

Kelly, R.T. (2020). Single-cell proteomics: Progress and prospects. Mol. Cell Proteomics *19*, 1739–1748.

Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.

Kibinge, N., Relton, C., Gaunt, T., and Richardson, T. (2020). Characterizing the causal pathway for genetic variants associated with neurological phenotypes using human brain-derived proteome data. Am J Hum Genet *106*, 885–892.

Kiviet, D.J., Nghe, P., Walker, N., Boulineau, S., Sunderlikova, V., and Tans, S.J. (2014). Stochasticity of metabolism and growth at the single-cell level. Nature *514*, 376–379.

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. *19*, 1454–1462.

Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., et al. (2020). Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. Cell *183*, 1436–1456 e31.

Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol *20*, 681–697.

Li, J., Li, X., Zhang, S., and Snyder, M. (2019). Gene-environment interaction in the era of precision medicine. Cell *177*, 38–44.

Li, J., Ma, Z., Shi, M., Malty, R.H., Aoki, H., Minic, Z., Phanse, S., Jin, K., Wall, D.P., Zhang, Z., et al. (2015). Identification of human neuronal protein complexes reveals biochemical activities and convergent mechanisms of action in autism spectrum disorders. Cell Syst. *1*, 361–374.

Li, J., Pan, C., Zhang, S., Spin, J.M., Deng, A., Leung, L.L.K., Dalman, R.L., Tsao, P.S., and

Snyder, M. (2018). Decoding the genomics of abdominal aortic aneurysm. Cell *174*, 1361–1372 e10.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell *165*, 535–550.

Malty, R.H., Aoki, H., Kumar, A., Phanse, S., Amin, S., Zhang, Q., Minic, Z., Goebels, F., Musso, G., Wu, Z., et al. (2017). A map of human mitochondrial protein interactions linked to neurodegeneration reveals new mechanisms of redox homeostasis and NF-kappaB signaling. Cell Syst. *5*, 564–577 e12.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., Mccarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Medgadget. (2020). MIT's comfortable shirts loaded with body sensors. https://www.medgadget.com/2020/04/mits-comfortable-shirts-loaded-with-body-sensors.html.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., and Barabasi, A.L. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science *347*, 1257601.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

Mishra, T., Wang, M., Metwally, A.A., Bogu, G.K., Brooks, A.W., Bahmani, A., Alavi, A., Celli, A., Higgs, E., Dagan-Rosenfeld, O., et al. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. Nat. Biomed. Eng. *4*, 1208–1220.

Morel, D., Yu, K.C., Liu-Ferrara, A., Caceres-Suriel, A.J., Kurtz, S.G., and Tabak, Y.P. (2020). Predicting hospital readmission in patients with mental or substance use disorders: a machine learning approach. Int. J. Med. Inform *139*, 104136.

Moutaoufik, M.T., Malty, R., Amin, S., Zhang, Q., Phanse, S., Gagarinova, A., Zilocchi, M., Hoell, L., Minic, Z., Gagarinova, M., et al. (2019). Rewiring of the human mitochondrial interactome during neuronal reprogramming reveals regulators of the respirasome and neurogenesis. iScience *19*, 1114–1132.

Nesterova, A.P., Yuryev, A., Klimov, E.A., Zharkova, M., Shkrob, M., Ivanikova, N.V., Sozin, S., and Sobolev, V. (2019). Disease Pathways : An Atlas of Human Disease Signaling Pathways (Elsevier).

Quer, G., Radin, J.M., Gadaleta, M., Baca-Motes, K., Ariniello, L., Ramos, E., Kheterpal, V., Topol, E.J., and Steinhubl, S.R. (2021). Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat. Med. *27*, 73–77.

Robins, C., Liu, Y., Fan, W., Duong, D.M., Meigs, J., Harerimana, N.V., Gerasimov, E.S., Dammer, E.B., Cutler, D.J., Beach, T.G., et al. (2021). Genetic control of the human brain proteome. Am. J. Hum. Genet. *108*, 400–410.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.

Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018). Genomic atlas of the human plasma proteome. Nature *558*, 73–79.

Sund-Levander, M., Forsberg, C., and Wahren, L.K. (2002). Normal oral, rectal, tympanic and axillary body temperature in adult men and women: a systematic literature review. Scand. J. Caring Sci. *16*, 122–128.

Uddin, M., Wang, Y., and Woodbury-Smith, M. (2019). Artificial intelligence for precision medicine in neurodevelopmental disorders. NPJ Digit. Med. *2*, 112.

Wang, C., and Li, J. (2020). A deep learning framework identifies pathogenic noncoding somatic mutations from personal prostate cancer genomes. Cancer Res. *80*, 4644–4654.

Wang, D., Cui, P. & Zhu, W. 2016. Structural Deep Network Embedding. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1225-1234.

Wierbowski, S.D., Liang, S., Chen, Y., Andre, N.M., Lipkin, S.M., Whittaker, G.R., and Yu, H. (2020). A 3D structural interactome to explore the impact of evolutionary divergence, population variation, and small-molecule drugs on SARS-CoV-2-human protein-protein interactions. bioRxiv, 2020.10.13.308676.

Willsey, A.J., Morris, M.T., Wang, S., Willsey, H.R., Sun, N., Teerikorpi, N., Baum, T.B., Cagney, G., Bender, K.J., Desai, T.A., et al. (2018). The psychiatric cell map initiative: a convergent systems biological approach to illuminating key molecular pathways in neuropsychiatric disorders. Cell *174*, 505–520.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature *499*, 79–82.

Wu, P., Heins, Z.J., Muller, J.T., Katsnelson, L., De Bruijn, I., Abeshouse, A.A., Schultz, N., Fenyo, D., and Gao, J. (2019). Integration and analysis of CPTAC proteomics data in the context of cancer genomics in the cBioPortal. Mol. Cell Proteomics *18*, 1893–1898.

Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B.B., Laser, A., Maranville, J.C., Wu, H., et al. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat. Commun. *9*, 3268.

Ye, Y., Zhang, Z., Liu, Y., Diao, L., and Han, L. (2020). A multi-omics perspective of quantitative trait loci in precision medicine. Trends Genet. *36*, 318–336.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhou, X., Menche, J., Barabasi, A.L., and Sharma, A. (2014). Human symptoms-disease network. Nat. Commun. *5*, 4212.

Zhu, Y., Piehowski, P.D., Zhao, R., Chen, J., Shen, Y., Moore, R.J., Shukla, A.K., Petyuk, V.A., Campbell-Thompson, M., Mathews, C.E., et al. (2018). Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. Nat. Commun. *9*, 882.

Zilocchi, M., Moutaoufik, M.T., Jessulat, M., Phanse, S., Aly, K.A., and Babu, M. (2020). Misconnecting the dots: altered mitochondrial protein-protein interactions and their role in neurodegenerative disorders. Expert Rev. Proteomics *17*, 119–136.