π-HuB: The Proteomic Navigator of The Human Body

Fuchu He[1,2]*, Ruedi Aebersold[3]*, Mark S. Baker[4], Xiuwu Bian[5], Xiaochen Bo[6], Daniel W Chan[7], Cheng Chang[1], Luonan Chen[8], Xiangmei Chen[9], Yu-Ju Chen[10], Heping Cheng[11], Ben C Collins[12], Fernando Corrales[13], Jürgen Cox[14], Weinan E[15], Jennifer Van Eyk[16], Jia Fan[17], Pouya Faridi[18], Daniel Figeys[19], George Fu Gao[20], Wen Gao[21], Zu-Hua Gao[22], Keisuke Goda[23], Wilson Wen Bin Goh[24], Dongfeng Gu[25], Changjiang Guo[26], Tiannan Guo[27], Yuezhong He[2], Albert J.R. Heck[28], Henning Hermjakob[29], Tony Hunter[30], Narayanan Gopalakrishna Iyer[31], Ying Jiang[1], Connie R Jimenez[32], Lokesh Joshi[33], Neil L Kelleher[34], Ming Li[35], Yang Li[1], Qingsong Lin[36], Cui Hua Liu[37], Fan Liu[38], Guang-Hui Liu[39], Yansheng Liu[40], Zhihua Liu[41], Teck Yew Low[42], Ben Lu[43], Matthias Mann[44], Anming Meng[45], Robert L. Moritz[46], Edouard Nice[47], Guang Ning[48], Gilbert S Omenn[49], Christopher M Overall[50], Giuseppe Palmisano[51], Yaojin Peng[52], Charles Pineau[53], Terence Chuen Wai Poon[54], Anthony Purcell[55], Jie Qiao[56], Roger Reddel[57], Phil Robinson[57], Paola Roncada[58], Chris Sander[59], Jiahao Sha[60], Erwei Song[61], Sanjeeva Srivastava[62], Aihua Sun[1], Siu Kwan Sze[63], Chao Tang[64], Liujun Tang[1], Ruijun Tian[65], Juan Antonio Vizcaíno[29], Chanjuan Wang[1], Chen Wang[66], Xiaowen Wang[1], Xinxing Wang[26], Yan Wang[1], Tobias Weiss[67], Mathias Wilhelm[68], Robert Winkler[69], Bernd Wollscheid[70], Limsoon Wong[71], Linhai Xie[1], Wei Xie[45], Tao Xu[72,73], Tianhao Xu[2], Liying Yan[56], Jing Yang[72], Xiao Yang[1], John Yates[74], Tao Yun[75], Qiwei Zhai[76], Bing Zhang[77], Hui Zhang[78], Lihua Zhang[79], Lingqiang Zhang[1], Pingwen Zhang[80], Yukui Zhang[79], Yu Zi Zheng[22], Qing Zhong[57], Yunping Zhu[1] & The π-HuB Consortium[81]

A list of affiliations appears at the end of the paper.

Abstract

The human body contains trillions of cells, classified in specific cell types, diverse morphologies and functions. Additionally, cells of the same type can assume different states within an individual's body during their lifetime. Understanding the complexities of the proteome in the context of a human organism and its many potential states is a necessary requirement to understand human biology, but these complexities can neither be predicted from the genome, nor have they been systematically measurable with available technologies. Recent advances in proteomic technology and computational sciences now provide opportunities to investigate the intricate biology of the human body at unprecedented resolution and scale. Here, we introduce a big-science endeavor called π-HuB (**P**roteomic Nav**i**gator of the **Hu**man **B**ody). The ultimate aim of the π-HuB project is to generate and harness multi-modality proteomic datasets to enhance our understanding of human biology, to facilitate disease risk assessment, diagnosis, to uncover new drug targets, to optimize appropriate therapeutic strategies, and to enable intelligent healthcare, thereby ushering in a new era of proteomics-driven phronesis medicine. This ambitious mission will be implemented by an international collaborative force of multidisciplinary research teams worldwide across academic, industrial and government sectors.

Introduction

The Human Genome Project (HGP) provided a comprehensive map of the human genome and identified the specie's ~20,300 protein coding genes[1,2]. It demonstrated the power of data-driven, large-scale coordinated 'omics' projects in transforming biomedical research, giving rise to genomics-driven precision medicine (GDPM). The human body contains approximately 37 trillion cells of distinctive types, diverse morphologies and functions, organized in tissues and organs, all of which share essentially the same genome. Moreover, during individuals' lifetime, tissues/organs and cells within their bodies have often undergone extensive or reversible/irreversible changes in response to changing conditions. The experience over the approximately 24 years since the publication of the human genome sequence has shown that the observed cellular and organismic complexities cannot be predicted from the genomic information alone.

Human biology, with all its intricate complexities, is profoundly interconnected with the vast expanse that is often termed the protein 'universe' or, more scientifically, the proteome (the complete set of proteins expressed by a genome in any cell or tissue at a specific point in time)[3]. Serving as the fundamental functional elements of cellular mechanisms, proteins are involved in essentially any biological process within an organism. Their significance extends beyond normal physiology; proteins have been crucially implicated as major contributors to the onset and progression of various diseases. They emerge as central figures in the field of therapeutics, being the primary molecular targets for a large majority of drugs. Hence, far beyond the static view provided by genomics, proteomics provides information about the dynamic aspects of the human body and its adaption to changing conditions. Following the HGP era, research based on the human proteome is one of the most exciting, yet challenging topics in life sciences and medicine[4,5].

In 2001, coinciding with the publication of the human genome sequence, a group of proteomics researchers founded the International Human Proteome Organization (HUPO)[6]. In September 2010, HUPO took the first step towards an international collaborative effort, termed the Human Proteome Project (HPP), to find high-quality evidence for the expression of all human protein-coding genes using mass spectrometry (MS) and making them routinely and reliably measurable. Since then, HUPO has stimulated and coordinated many workshops to work in the HPP. Ten years later in 2020, the HUPO HPP project teams described the first high-stringency HPP proteome map, covering 90.4% of the canonical human proteome, paralleling similar decadal progress made by the HGP[7]. By 2023, detection of 18,397 (93%) of the predicted 19,750 canonical proteins encoded in the human genome had been achieved[8] and a compendium of validated reference spectra for the highly specific targeted mass spectrometric measurement >99% of the annotated human proteins had been generated[9]. Since the formation of HUPO, there has been an expansion of biology/disease-centric initiatives under the HPP umbrella. These aim to spatially measure and interpret human proteome data under a range of physiological and pathological conditions, including protein abundance, post-translational modifications, interaction partners, and localization.

The first proteomics project dedicated to a human organ (Human Liver Proteome Project, HLPP) was actually launched in China in 2003, much before, and as a forerunner of, the HPP[10-12]. This led to the characterization of liver protein expression profiles and protein-protein interactions in this metabolic organ[13], as well as the discovery of a major functional role of acetylation in metabolic regulation[14,15]. Subsequently, the proteomes of other tissues or organs (e.g., brain[16], heart[17], stomach[18], skin[19], and immune cells[20]) have been characterized, creating an initial version of the organ/tissue-based human proteome map[21,22]. Meanwhile, an increasing number of disease-related organ/tissue proteomes have been analyzed as exemplified by the Chinese Human Proteome Project (CNHPP)[23], the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) in the United States[24], the Tumor Profiler Project (TuPro) in Switzerland[25], the Human Protein Atlas in Sweden[26,27], and ProCan in Australia[28]. Moreover, recent advances in non-MS based approaches provide versatile opportunities for biomarker discovery in body fluids, which is thought to reflect a person's health or disease status[29-31]. All these efforts are providing a significant boost in advancing the field toward an era that has previously been termed PDPM (proteomics-driven precision medicine)[32]. Despite remarkable technological and computational advances, we are only just beginning the exploration of the complexities of the human proteome, and the exploitation of the full potential of the proteome for biomedical breakthroughs has yet to be fully harnessed (**Box 1**).

In 2020, the Chinese Ministry of Science and Technology funded a collaboration of around 40 proteomics research teams worldwide to envision future HPP-related projects. Since then, several multidisciplinary working groups have been established, and numerous on-site meetings and webinars have been organized, which have been communicated with government and private funding bodies. These activities have allowed us to propose a big science project called π-HuB (Proteomic Navigator of the Human Body). The project is forming a consortium of Chinese and international scientists to generate mega proteomic datasets from all major human tissues/organs and cell types, and to subject the data to integrative analysis at an unprecedented scale. The ultimate aim is to build an intelligent computational engine called the π-HuB navigator which will integrate multi-modality proteomic datasets to enhance our understanding of human biology, to facilitate disease risk assessment, diagnosis, to uncover new drug targets, to optimize appropriate therapeutic strategies, and to enable intelligent healthcare.

Three Central Goals of the π-HuB Project

The π-HuB project has the overriding mission, to support, with an investment of billions of RMB, the international and interdisciplinary collaboration of scientists for a period of three decades towards three specific goals (**Fig. 1**).

1. Discover principles of the human body The π-HuB project will first dissect the human body into a hierarchy of digital proteomic anatomy spaces. By harnessing rapidly evolving techniques such as single-

cell and spatial proteomics[33,34], the project will digitize, and continually refine, the complete quantitative proteomic and cellular composition of the human body, including cell composition of all major tissues/organs, protein composition of individual cell types and single cells, and proteome-centric molecular networks within cells (e.g., PTMs, PPIs). Taking advantage of recent advances in multi-modal data fusion/integration technology, in particular, the rapid development of deep learning or foundational models, the high-resolution anatomy-based proteomic data will provide an unprecedented opportunity to decode the essential molecular/cellular building principles of cells/tissues/organs, and to uncover the critical molecular/cellular mechanisms of biological processes, i.e. to reveal causal relations from a protein network to a phenotype.

2. Develop the Meta Homo Sapiens model The π-HuB project will conduct in-depth investigations into the dynamics of the human proteome throughout an individual's lifespan, exploring at a population level how the human proteome adapts in response to various factors impacting health outcomes. The entire human body state space will thus be transformed into multiple subspaces that are further dissected through various dimensions. Specifically, the aim is to trace proteome-centric trajectories during major prenatal and postnatal stages, to profile longitudinal dynamic changes in complex proteomes during the development and progression of representatives of complex diseases, and to determine the effects of non-genetic factors (e.g., symbiotic microbiomes, lifestyles, and different environments) on the human proteome. These state-contextual proteomic data will be integrated with other human omics data from complementary efforts (e.g., HuBMAP[35], Human Cell Atlas[36], Human Tumor Atlas Network[37], and the LifeTime Initiative[38]) and projected into a digitized model called *Meta Homo Sapiens*. Building such a model will be facilitated by *Composing Principles of the human body* and it will be formulated using a 3D anatomical hierarchy that records digital features of organs, tissues, body fluids and cells at each level and consist of time-sequential frames, with each containing proteomic data measured and augmented within a unit period to represent the human body state at a given timestamp.

3. Build the π-HuB Navigator The ultimate goal of the π-HuB project is to instantiate proteomics-driven phronesis medicine, which is a concept similar to the practical wisdom (Phrónesis) from ancient Greek that is relevant to practical action in a particular situation. Unlike traditional and current paradigms of medicine, phronesis medicine aims to develop the ability to provide temporally precise control of the human body state to prevent disease. This ability should include accurate, efficient monitoring, diagnosis and treatment capabilities and highly robust decision-making capabilities for disease prediction, early warning, prevention, control, and health care. It will ultimately provide temporally precise control of the human body state to prevent disease by establishing a medical model of popularization and normalization of monitoring, diagnosis and treatment decisions and health management. Undertaken with a very keen eye toward

realizing this ultimate goal, we aim to develop the π-HuB navigator that will be a virtual state-space instrument, created by the convergence of physiological phenotypes and proteomic-oriented spatial-temporal biochemical/biophysical information in cells, body fluids, tissues, and organs. It can transfer the prototype *Meta Homo Sapiens* model from primary body conditions to different secondary states to obtain realistic models. This will be followed by creating a state space covering all key states of the human body by simulating body dynamics with each model for specific periods, thereby tackling the most beneficial approaches to prediction of outcomes based on non-invasive proteomic snapshots and longitudinal proteomic measurements. Ultimately, causal inference will be used to identify underlying triggers that induce transitions between adjacent key states. Each state space can be regarded as a topological navigation map where each node is the key state defined by the corresponding biomarkers and each edge between two nodes records the triggers to transform from one state to the other. Thus, building such a navigator will grant an opportunity to track trajectories of wellness and health, to define factors important in disease risk assessment and early diagnosis, and to drive the development of novel therapeutic interventions and intelligent healthcare approaches for redirecting unhealthy transitions to a long and prosperous life.

Pillars for Building the π-HuB Navigator

To achieve the above goals, the project is being supported by six key pillars as follows (**Fig. 2**).

1. Human biospecimens Human biospecimens are the fundamental component of the π-HuB project. To achieve the objectives outlined above, samples for π-HuB can be grouped into the following categories. First, anatomy-based samples will consist of freshly prepared organs, tissues and live human samples obtained with the highest ethical standards from post-mortem examinations. Second, twin cohorts will allow the calculation of the genetic component of observed variability in a population and benefit on controlling confounding factors in etiological studies on complex diseases. Third, population-based cohorts will be cross-sectional collections of high-quality biospecimens from a large number of individuals from diverse geographical regions of the world with different lifestyles and subjected to different environments. Finally, longitudinal cohort studies will apply non- or less-invasive approaches with relatively high sampling frequency to sample healthy individuals or patients with defined exposures that have health or therapeutic implications or outcomes.

In practice, the π-HuB project will first utilize samples from existing state-of-the-art biobanks around the world and is open to work closely with other resources being built. All samples in this project will be required to be well and consistently annotated with clinical and demographic information obtained from multiple sources, such as questionnaires, physical measurements, biochemical tests, medical imaging data, records of genetic variants implicated in disease susceptibility, and wearable device-based records, among others. Furthermore, annotations should be performed by using agreed metadata standards that are key for

data accessibility and interoperability and for artificial intelligence (AI)-based data integration across bioinformatic resources[39].

2. Measurement technology innovations Considering that the project aims to profile the human proteome at unprecedented resolution and scale and that it has an intended time horizon of 30 years, the advancement of measurement technologies is crucial to its success (**Box 2**). For example, there is a pressing need to identify and measure minute amounts of proteins from single cells. In recent years, we have witnessed tremendous progress in the development of MS-based single-cell proteomic (SCP) technologies[40]. However, a real-world, large-scale application of existing technologies for profiling millions to billions of human cells, is still very far from the state-of-the-art. Whereas current MS-based SCP technologies are able to measure up to ~4,500 proteins per single cell[41], the limited sample throughput due to the lack of multiplexing strategies remains a major shortcoming of SCP analyses. In this regard, the project will start by profiling a cell type-resolved human proteome atlas (See below), while in parallel π-HuB will coordinate and support an international community effort to accelerate and benchmark MS-based SCP technologies across different platforms and laboratories. Once we can achieve a relatively high analytical performance for SCP analysis in the π-HuB data collection centers (e.g., >3,000 proteins per cell at a throughput of ~1,000 cells per day), the project will then launch the first initiative to collect SCP data from human samples. Meanwhile, we will closely track novel concepts and technologies for single-molecule protein sequencing that have substantial potential to enable broad sequence coverage in single-cell profiling[42].

In addition, the π-HuB project will also develop, integrate and apply robust technologies for the generation of multidimensional proteomic data indicating the functional states of the proteome. These technologies are exemplified by PTMs, structural states, localization, and interactions and context-specific interdependencies of these attributes. To facilitate this, the π-HuB project will develop a specific technology scout division separate from its own technology hub, seeking to identify and fund emerging technologies.

As technologies will be rapidly evolving, data acquired in the early stage of this project may become superseded by data collected at a later stage. However, these early data will be invaluable for the development and benchmarking of data analysis and integration tools, for providing training opportunities for researchers in the field, for demonstrating data and sample interoperability procedures within the consortium and for supporting pilot studies. Moreover, we reasoned that, like HGP, the project will also proceed in stages where 'drafts' of the high-fidelity human proteome atlases will become available at regular intervals, which will be further iterated to more accurate and complete versions by using newer technologies.

3. Computational technology innovations Beyond data collection, our ambition also extends to developing methods and tools for data integration, analysis and interpretation (**Box 2**). Data-driven modeling approaches, such as automated machine learning (autoML), have proved powerful in approximating many

virtual and real-world systems[43]. However, transforming a biological 'black box' into a digital system does not typically provide us with any intellectual knowledge or insight that would enable it to be trusted for clinical practice. Therefore, the π-HuB project will push the boundaries in biomedicine by unveiling the molecular re-construction of the human body. Inspired by success in mathematical intuition guidance and hypothesis proposal[44], explainable artificial intelligence (XAI) methods, large language models (LLMs)[45], and other yet-to-be-conceived approaches will be exploited to interpret a fit-for-purpose deep learning model of the human body with escalating resolutions, from molecular to cellular to organ and systems levels, enabling the discovery of knowledge about biological events and the establishment of construction theories[46]. In this context, it will be key to attract AI practitioners to the proteomics field. With this new knowledge, several 'white box' prototype *meta-Homo sapiens* models will be constructed to serve as the critical preliminary conditions upon which π-HuB will build the foundation model of the system. We believe that the π-HuB project is ideally positioned to meet the data science challenges because of the availability of proteomic data sets acquired by the project consortium that are unique in terms of the size, consistency of collection, annotation and processing and their coverage of multiple layers of the proteome.

4. Big-Science infrastructure Since very large numbers of human samples (e.g., sorted single cells from human organs and biospecimens from clinical cohorts) will be analyzed in the π-HuB project, ultra-high throughput facilities for data manufacturing, collection and processing will be required. Ideally, such facilities require expertise and streamlined, reproducible pipelines to process human samples, profile proteome-centric molecular data in samples, and store, transfer, process, and interpret those data. Therefore, the π-HuB project will be establishing national facilities/centers as the Big Science infrastructure for the collection and processing of multi-layer proteomic data. Minimally, during the first stage of the π-HuB project, such an infrastructure should be able to process 1,000-2,000 samples per day (SPD) and generate 1TB of MS raw data per day (as of today, although this amount is expected to increase in parallel with the developments in technology and instrumentation). In China, only a few existing programs possess such analytical capacities including automation workstations for 'one-stop' sample preparation, more than 40 cutting-edge high-resolution mass spectrometers and a high-performance computing system called 'Tianhe-II'[23]. In addition, many other Big Science infrastructures and National Laboratories across China have pledged support for the π-HuB project, bringing state-of-the-art single cell technologies, multimode trans-scale biomedical imaging technologies and a cloud-based high-performance AI computing system to the project. Furthermore, the π-HuB project is partnering with existing infrastructures from research entities attached to universities or other institutions worldwide, such as the Netherlands Proteomics Center and ProCan in Australia.

5. Open resources The π-HuB project will emphasize highly efficient, international, open resources, including standards, samples and their annotations, data, and key analysis tools. Like other large community

resources with a broad utility, the project will require an open sharing framework to ensure transparent global collaboration between researchers, funding agencies, and stakeholders. In this framework, π-HuB will maximize the importance of reusing collected human samples and reanalyzing already generated data to maximize the benefits from scientific advances, while minimizing risks to participant privacy and acknowledging the contributions of researchers. For example, all π-HuB-generated (non-sensitive) raw data will be directly available to the international scientific community through several well-established data portals implementing the FAIR (Findable, Accessible, Interoperable and Reusable) data principles, such as those established by the ProteomeXchange Consortium, including PRIDE and iProX[47,48]. In addition, bioinformatics infrastructure will be developed to integrate proteomics atlases into UniProt, the most popular protein knowledgebase in the world, so that information is available to the whole life science community. Moreover, the project will enable clinicians and patients to freely inquire about medical intervention strategies by developing a web-based '*Meta Homo sapiens*' computational framework based on π-HuB molecular and spatial data.

6. International research teams The sixth pillar is 'people', including researchers, software engineers, clinicians, pathologists, project managers, administrators, financial staff, lawyers, commercial entities and so on. The implementation of the π-HuB project requires synergy between a plethora of people globally to work collaboratively under the guidance of a decision-making body and clear governance and accountability guidelines. Specifically, the π-HuB project will be steered and governed by an Executive Committee and overseen an advisory board. In addition, capacity building and cultural interchange will benefit international researchers in the exchange of ideas and results and also in research and intellectual culture. To foster this, π-HuB will establish several scholarship or fellowship programs to promote these exchanges, through which it will attract additional early career scientists to participate in this visionary international project.

Challenges

1. Ethics Within the framework of the π-HuB project, several measures are proposed to mitigate the identified ethical and regulatory challenges in proteomic research[49]. Especially with regards to human samples, it is essential for the project to establish a common, flexible and generally accepted framework with respect to ethics approval criteria, patient/donor consent, sample annotation ontology material transfer agreements (MTA) and non-disclosure agreements (NDA) that can be accepted by governments from different geographical regions. Additionally, the risk of re-identifying individuals through their proteomic features demands meticulous assessment and management. To safeguard the ethical integrity and foster societal acceptance of the π-HuB project, we will establish a specialized ethics committee whose primary role will be to oversee every facet of the project, encompassing biospecimen collection, analysis, data

management, and information dissemination. Adopting this proactive stance towards ethical and regulatory compliance will not only enhance the scientific credibility of the π-HuB project but also strengthen public trust and participation.

2. Big Data π-HuB will work closely with existing international data centers such as the ProteomeXchange consortium for the sake of consistency of data standards and management rules widely adopted in the field of proteomics. Nonetheless, a fit-for-purpose data center specifically designed for the π-HuB project is needed. First, new upgrades will be required to store and manage more metadata (e.g., clinical information) that can meet the data management requirements of the project. Second, it should support multi-omics and multi-model data management and applications. Last but not least, it needs to be easily accessible for researchers worldwide. This will be achieved by the establishment of sub-centers in different countries that can meet the legal constraints of each country[50], while meeting the requirements of the π-HuB project by using data management system software and technical support from the headquarter center.

3. Data generation and integration The massive amount of data from different modalities will be collected across the international teams. Thus, it is necessary to ensure that data generated across different teams are comparable and integrable. The π-HuB consortium will first ensure all human samples to be processed using standard operating procedures (SOPs) with respect to collection, annotation, handling, storage, and tracking. Additionally, to ensure that each data collection team can produce high-quality and unified proteome-centric datasets, the π-HuB project will adopt the HUPO Proteome Standards Initiative (PSI)[51] principles to standardize both state-of-the-art MS-based and non-MS-based approaches in terms of sample preparation, methodological settings, data acquisition, processing and error control, and develop the SOPs for each step. Given the anticipated rapidity of methodological development, the consortium will also develop and share standards, test samples, and benchmarking data to help each research center update SOPs for new techniques.

Furthermore, new computational methods and ML models with strong generalization ability[52] are needed to further develop proteomic data analysis (e.g., quality control[53], data cleansing, normalization[54], and missing-value imputation) and specifically address questions that can benefit from multimodal and inter-center/laboratory measurement. As such, we will develop a centralized, cloud-based, interactive platform for data sharing and analysis that will host standardized tools and pipelines for data processing, integration and interpretation. Lastly, the consortium will provide comprehensive training and support to all consortium members to ensure familiarity and compliance with the SOPs and computational tools. These efforts will develop approaches to support the mobilization of the π-HuB data, support the discovery of novel insights by providing novel algorithms and develop novel models for unifying multiple-omics layers.

4. Modeling The π-HuB Navigator will build upon a computationally driven model (so called *Meta Homo Sapiens*) of the human proteome, which is an extremely complex task. In our initial plan, this model will

consist of three basic modules (**Fig. 3**): 1) a state identifier to encode different states of the human body in the state space through proteome-centric measurements, followed by the integration of phenotypic information of the human body through a multi-modality LLM. 2) a lineage tracer to quantify the transition probability between each pair of states under different physiological/pathological/therapeutic conditions; for example, the transition probability between each pair of states can be estimated through Monte Carlo methods[55]. 3) a path planner to search for the optimal treatment trajectory by balancing various objectives such as the efficacy against the financial costs and individual's compliance.

5. Democratizing proteomics Compared to genomics and its related sequencing technologies, the power of proteomics is far from being fully appreciated by the public and, in fact many clinicians. Thus, building the interface of the π-HuB project to clinicians and the public is a major aim in addition to the research goals of the project to gain public awareness and participation. Such advocacy is needed as a driving force in both sample procurement and addressing the most impactful and pressing needs in disease focused research. The π-HuB consortium will also provide training and education of clinicians, pathologists, and patients to interpret and use proteomics data, and push proteomics-driven discoveries to the clinics and health care. In particular, the tools hosted in this interface will assist researchers and clinicians in understanding the biological pathways of specific state changes, guiding daily research and clinical practice.

Major Outcomes of π-HuB Phase One

To enable π-HuB to be a broadly applicable project, it is necessary to maximize relevance to the community by setting deliverables and expected outcomes as a series of staged programs undertaken in a relatively short timeframe. During the initiation and development stage (2024-2033, herein referred to as 'phase one'), we will build an international cooperative network to lay the technical foundation of this project by promoting methodological advances, benchmarking state-of-the-art technologies for standardization, building the computational infrastructures for data integration and modeling, and so on. Meanwhile, it is also important for the project, in a relatively short-term frame, to achieve major outcomes as follows (**Table 1**).

1. Principles of cell type organization The π-HuB project will eventually support studies that generate single cell resolution atlases of all major human organs and tissues from people who identify as emanating from different ancestral populations/origin. However, during phase one, we will initially build reference cell-type proteome atlases for all major organs using a combination of state-of-the-art flow sorting[56] and parallel MS acquisition platforms[57]. Moreover, rapidly evolving spatial proteomics technologies will provide additional insights into the secreted proteins in the surrounding microenvironment and subcellular localization of the proteome at the tissue, cellular, and molecular levels. In this regard, multiple cutting-edge and synergistic approaches will be implemented, including MS-based, multiplex immuno-affinity-

based and super-resolution imaging-based methods[58,59]. Additionally, the emerging proximity labeling and *in vivo* crosslinking approaches will enable the profiling of protein subcellular localization, protein complexes and PPIs in diverse human cell types[60,61]. These analyses will provide versatile opportunities to uncover new molecular/cellular mechanisms of biological processes in shaping diverse cell types and cell states within each organ. Together, we envision that the cell-type-resolved, multidimensional proteome atlases, in combination with cutting-edge computational and bioinformatic approaches, are able to uncover building principles of cell type organization of major tissues/organs.

2. Proteomics-driven lifestyle guidelines During phase one, we will focus on the most dominant factors that shape/remodel the proteome of healthy individuals. Specifically, we will accumulate a large number of biofluid proteomes from large-scale natural populations, aiming 1) to map quantitative trait loci for circulating/tissue proteins and protein allelic variants associated with genetic variants implicated in disease susceptibility, allowing us to construct disease-causing pathways; 2) to trace the proteomic trajectory over the human lifespan by quantifying dynamic changes in the biofluid proteomes of five major prenatal cycles (e.g., gametogenesis, fertilization, embryonic development, fetal development, and delivery) and five major postnatal cycles (e.g., adolescence, puberty, gestation, menopause and old age); 3) to analyze the effects of four major dietary nutrition patterns (i.e., Western, Japanese, Mediterranean and subsistence)[62] on the human biofluid proteome; 4) to map the proteomes of populations in six major ecological environments that are classified by the Köppen-Geiger map (e.g., hot, warm, cold, arid, polar and highland)[63], and analyze the trajectory of the human biofluid proteome during acclimatization and adaptation; 5) to map interactions of the human gut and skin proteome with representative microbiomes from internal and external environments, and to construct the adaptation trajectory of the human proteome in response to microbiomes; and 6) to map the responses of the human proteome to various clinical intervention strategies such as medication, diet, and exercise. Together, these analyses will generate a resource of human proteome traits associated with the lifetime states as indicated above. Such a resource will provide opportunities to develop a proteomic health score through a neural network or LLM, thereby shaping up proteomics-driven lifestyle guidelines.

3. Generalization of PDPM In the past decade, there has been growing evidence that proteomic approaches can facilitate the mechanistic understanding of diseases as well as facilitate biomarker discovery and optimize therapy development. In particular, proteomics alone has been able to identify potential biomarkers or potential therapeutic targets for many tumor types[32,64-67] and a variety of other diseases[68-73]. Despite these advances, most proteomic findings in the context of human diseases have yet to be validated and treatment suggestions arising from the data have yet to be approved. For example, most potential biomarkers identified by proteomic studies are generated from small-scale, retrospective studies, lacking the basis for the subsequent generalization of using such biomarkers in a wider population. The π-HuB

consortium that consists of multi-interdisciplinary researchers and clinicians provides an unprecedented opportunity to reimagine biomarker discovery through a proteomic approach. We therefore reason that, under the umbrella of π-HuB, it will be more feasible to organize large-scale, international, multicenter, cohort studies for validating new biomarkers for early and companion diagnosis of major diseases. As such, we plan to map the proteomes of 10 major organs and corresponding biofluids at different pathophysiological stages, focusing on 3-5 representative diseases for each related organ. Such analyses, together with the aforementioned tissue proteome atlases with cell type resolution and biofluid proteome atlases with life-oriented adaptive proteome atlases, will allow the construction of proteomic evolutionary trajectories mapping the occurrence and development of these diseases and the pathways associated with specific life stages and survival conditions. Furthermore, the π-HuB project will actively collaborate with clinicians, policymakers, and industrial partners to catalyze the discovery of new protein-based biomarkers and drug targets that can be applied in clinics to diagnose disease and drug development, driving a paradigm shift to proteomics-driven precision medicine.

Outlook

Since its inception in 2020, the π-HuB consortium has grown to be an international collaborative force of more than 100 members mobilizing scientists worldwide across academic, industrial and government sectors in the protein and health sciences. The π-HuB project will likely foster further global collaboration and discussion by integrating the results of a worldwide community of multidisciplinary scientists to better understand human biology and to advance medicine from disease trajectory predictions to new treatment options. We anticipate that the π-HuB project will make a major contribution to biomedical research in the coming decades, facilitating disease prevention and diagnosis, accelerating drug discovery and ultimately ushering in an era of proteomics-driven phronesis medicine.

Box 1. Complexities of The Human Proteome

First, for a human being at any time, there is immense molecular diversity of the proteins in the human proteome at multiple scales, including their level of expression and degradation, their functional state indicated by post-translational modifications (PTMs), protein-protein interactions (PPIs) and shapes and their cellular and subcellular location. Collectively the proteins and their attributes shape the approximately 37 trillion cells in the human body with a wide range of morphology and function. Second, for any human society, the genomic diversity of the human population leads to a larger diversity of proteomes in the population, as each person will have a special private proteome and therefore a special private functional state. Third, during human lifetime, an individual's proteome is highly dynamic and can be affected by disparate external and internal factors, such as somatic mutations, the human microbiome (which can be defined as the microbial ecosystems that reside in various habitats of the body, for example, the human gut), the type of lifestyle (e.g., diet, food, nutritional supplements, physical activity and drugs), the occurrence of somatic mutations and the state of the external environment, all of which are intimately related to human health and diseases.

Box 2. Key Technologies for the π-HuB Project

- **Single Cell Proteomics:** π-HuB will fully benchmark state-of-the-art SCP methods (e.g., nanoPOTS[74], SCoPE-MS[75], and scPiMS[76]) and determine the right time to launch the first large-scale initiative to collect SCP data from human samples. In order to further increase the throughput of SCP analysis, it is also important to urge engineering-level innovations for fully integrated/automated ng-level sample preparation technologies[77] and single-molecule protein sequencing technologies[78].
- **Spatial Proteomics:** π-HuB will initially apply the Deep Visual Proteomics (DVP) technology[79] or its derivatives to spatially profile proteomes across different cell types. Nonetheless, new concepts and technologies for spatial proteomics are warranted by integrating AI-based tissue imaging navigation, high-throughput and pixel-format sampling, multimodal data acquisition and integration[34].
- **Plasma Proteome Profiling:** π-HuB will apply MS-based or affinity-based technologies for plasma proteome profiling, which have been demonstrated to simultaneously analyze thousands of proteins in many thousands of plasma samples with high throughput[80,81].
- **Functional Proteomics:** π-HuB will focus on new chemical biological and biophysical approaches for targeting and enriching native functional states of the human proteome. For example, recent technological advances have enabled the direct detection of subcellular localization, dynamic changes, and interactions of proteins *in vivo*[60,61,82,83].
- **Automated Machine Learning:** π-HuB aims to automate the end-to-end process of applying machine learning to analyze and interpret large-scale proteomics data[84,85], which involves tailoring the selection and optimization of machine learning models, facilitating non-expert access to complex multi-modal data analysis, and addressing challenges on the proteomics study such as protein identification, quantification, and biomarker discovery, early disease diagnosis, optimal therapeutic interventions, and dynamics of biological processes.
- **Explainable Artificial Intelligence:** π-HuB will develop XAI methods that provide clear and understandable explanations of their findings in proteomics analysis[86], which, in particular, can validate AI-driven hypotheses in proteomics, ensuring that AI conclusions are scientifically sound and interpretable, enhances trust and collaboration between computational scientists and experimental biologists by providing transparent decision-making processes. Furthermore, XAI can bridge the gap between AI models and practical applications (e.g., predictions of disease risk at the individual level and the effect of drugs or drug combinations on the state of cells).
- **Large Language Model:** π-HuB will build advanced AI models trained on extensive biomedical literature to understand and generate language specific to the proteomics field[45], which are expected to analyze and synthesize vast amounts of biomedical text, extract insights from unstructured data sources such as research papers relevant to proteomics, and assist in identifying patterns, potential therapeutic targets, and novel connections within the complex human proteomics data.

Figure legends

Figure 1. Overall goals of the π-HuB project. The schematic depicts the development and integration of major technological strategies for achieving the scientific milestones for each goal. The π-HuB project will start with the extensive measurement of human samples and the generation of data across digital proteomic anatomy spaces and individuals' state spaces in lifetime. Then, it will take the advantages of latest advances in data and computational sciences to uncover composing principles of the human body, to generate a digitized model called *Meta Homo Sapiens*, and to build a 'GPS' (Global Positioning System) for the human body and body's states.

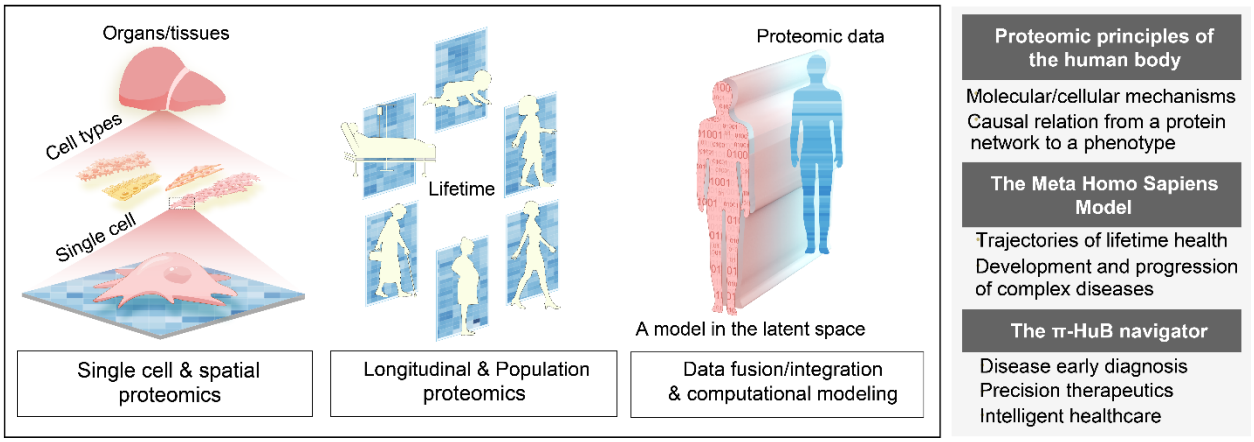


Figure 2. Key pillars for implementing the π-HuB project. To achieve the π-HuB project goals, the project is being supported by six pillars, of which key components are listed. autoML: Automated machine learning, XAI: Explainable artificial intelligence, LLM: large language model, HT: high-throughput, FAIR: Findable, Accessible, Interoperable and Reusable.

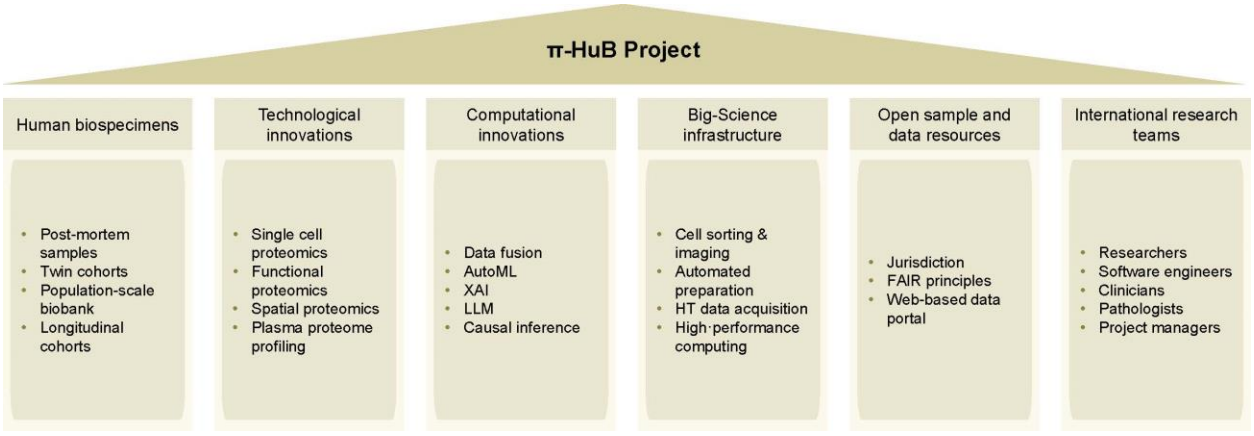
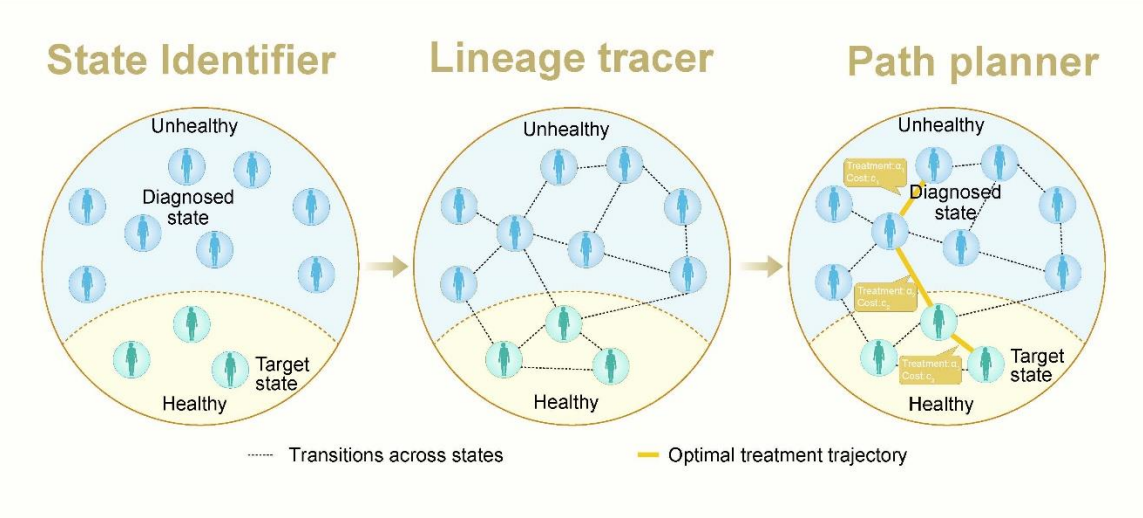


Figure 3. The basic modules of the π-HuB navigator. A state identifier generates key states of human body based on massive hierarchical measurements from large cohorts. It is followed by a dynamic model-based lineage tracer to detect possible transitions between states under all available treatments. Given the

established state space, a path planner can search for a treatment trajectory while considering various objectives and constraints.



Table

Table 1. Major outcomes of π-HuB Phase One

Expected Outcomes	Principles of cell type-based tissue organization	Proteomics-driven lifestyle guidelines	Proteomics-driven precision medicine
Biospecimen inputs	All major organs from post-mortem of healthy donors	Biobanks of natural populaiton	Large-scale international multicenter patient cohorts
Key measurements	Protein expression, subcellular localization, PTMs and PPIs in each cell type	Protein expression and PTMs in body fluids or other non-invasive human samples	Protein expression, PTMs, and PPIs indisease tissues and/or body fluids
Major deliverables	1. A cell-type-resolved, multidimensional human proteome atlas 2. New molecular/cellular mechanisms of biological processes	1. A resource of human proteome traits associated with lifetime states 2. A proteomic health score	1. A resource of human proteome traits associated with major diseases 2. New biomarkers and therapeutic targets

References

1 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

2 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

3 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355 (2016).

4 Abbott, A. And now for the proteome. *Nature* **409**, 747 (2001).

5 Fields, S. Proteomics. Proteomics in genomeland. *Science* **291**, 1221-1224 (2001).

6 The proteome isn't genome II. *Nature* **410**, 725 (2001).

7 Adhikari, S. *et al.* A high-stringency blueprint of the human proteome. *Nat Commun* **11**, 5301 (2020).

8 Omenn, G. S. *et al.* The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. *J Proteome Res* **22**, 1024-1042 (2023).

9 Kusebauch, U. *et al.* Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **166**, 766-778 (2016).

10 Cyranoski, D. China takes centre stage for liver proteome. *Nature* **425**, 441 (2003).

11 Jia, H. & Louet, S. China pushes liver proteomics. *Nat Biotechnol* **22**, 136 (2004).

12 He, F. Human liver proteome project: plan, progress, and perspectives. *Mol Cell Proteomics* **4**, 1841-1848 (2005).

13 Wang, J. *et al.* Toward an understanding of the protein interaction network of the human liver. *Mol Syst Biol* **13**, 965 (2017).

14 Wang, Q. *et al.* Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science* **327**, 1004-1007 (2010).

15 Zhao, S. *et al.* Regulation of cellular metabolism by protein lysine acetylation. *Science* **327**, 1000-1004 (2010).

16 Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nat Neurosci* **18**, 1819-1831 (2015).

17 Doll, S. *et al.* Region and cell-type resolved quantitative proteomic map of the human heart. *Nat Commun* **8**, 1469 (2017).

18 Ni, X. *et al.* A region-resolved mucosa proteome of the human stomach. *Nat Commun* **10**, 39 (2019).

19 Dyring-Andersen, B. *et al.* Spatially and cell-type resolved quantitative proteomic atlas of healthy human skin. *Nat Commun* **11**, 5587 (2020).

20 Rieckmann, J. C. *et al.* Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol* **18**, 583-593 (2017).

21 Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582-587 (2014). **This paper, together with Ref22, that has revealed the initial version of the tissue/organ-centric human proteome by applying mass spectrometry-based approaches.**

22 Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509**, 575-581 (2014).

23 Cyranoski, D. China pushes for the proteome. *Nature* **467**, 380 (2010).

24 Rodriguez, H., Zenklusen, J. C., Staudt, L. M., Doroshow, J. H. & Lowy, D. R. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell* **184**, 1661-1670 (2021).

25 Irmisch, A. *et al.* The Tumor Profiler Study: integrated, multi-omic, functional tumor profiling for clinical decision support. *Cancer Cell* **39**, 288-293 (2021).

26 Uhlén M et al., **A human protein atlas for normal and cancer tissues based on antibody proteomics.** *Mol Cell Proteomics.* **4**, 1920-1932 (2005).

27 Uhlén M et al., **Tissue-based map of the human proteome.** *Science* **347**, 1260419 (2015).

28 Tully, B. *et al.* Addressing the Challenges of High-Throughput Cancer Tissue Proteomics for Clinical Application: ProCan. *Proteomics* **19**, e1900109 (2019).

29 Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348-358 (2023).

30 Xu, Y. *et al.* An atlas of genetic scores to predict multi-omic traits. *Nature* **616**, 123-131 (2023).

31 Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).

32 Jiang, Y. *et al.* Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **567**, 257-261 (2019). **The first study that has showcased the concept of proteomics-driven precision medicine.**

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

34 Mund, A., Brunner, A. D. & Mann, M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol*

Cell **82**, 2335-2349 (2022). **A review that has summarized state-of-the-art spatial proteomics for human samples.**
35 Hu, B. C. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187-192 (2019).
36 Elmentaite, R., Dominguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat Rev Genet* **23**, 395-410 (2022).
37 Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236-249 (2020).
38 Rajewsky, N. *et al.* LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377-386 (2020).
39 Mann, M., Kumar, C., Zeng, W. F. & Strauss, M. T. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst* **12**, 759-770 (2021).
40 Perkel, J. M. Single-cell proteomics takes centre stage. *Nature* **597**, 580-582 (2021).
41 Guzman, U. H. *et al.* Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat Biotechnol* (2024).
42 MacCoss, M. J. *et al.* Sampling the proteome by emerging single-molecule and mass spectrometry methods. *Nat Methods* **20**, 339-346 (2023). **A review that has summarized the next-generation proteomics technologies.**
43 Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533-538 (2023).
44 Davies, A. *et al.* Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70-74 (2021).
45 Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat Med* **29**, 1930-1940 (2023).
46 Kang, M., Ko, E. & Mersha, T. B. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform* **23** (2022).
47 Chen, T. *et al.* iProX in 2021: connecting proteomics data sharing with big data. *Nucleic Acids Res* **50**, D1522-D1527 (2022).
48 Deutsch, E. W. *et al.* The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res* **51**, D1539-D1548 (2023).
49 Fierro-Monti, I., Wright, J. C., Choudhary, J. S. & Vizcaino, J. A. Identifying individuals using proteomics: are we there yet? *Front Mol Biosci* **9**, 1062031 (2022).
50 Bandeira, N., Deutsch, E. W., Kohlbacher, O., Martens, L. & Vizcaino, J. A. Data Management of Sensitive Human Proteomics Data: Current Practices, Recommendations, and Perspectives for the Future. *Mol Cell Proteomics* **20**, 100071 (2021).
51 Deutsch, E. W. *et al.* Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work. *J Proteome Res* **22**, 287-301 (2023). **A perspective paper that has summarized the 20-year long community effort for the proteomics community with respects to data formats, quality control and annoation.**
52 Sharifi-Noghabi, H., Harjandi, P. A., Zolotareva, O., Collins, C. C. & Ester, M. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence* **3**, 962-972 (2021).
53 Olivella, R. *et al.* QCloud2: An Improved Cloud-based Quality-Control System for Mass-Spectrometry-based Proteomics Laboratories. *J Proteome Res* **20**, 2010-2013 (2021).
54 Chawade, A., Alexandersson, E. & Levander, F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res* **13**, 3114-3120 (2014).
55 James, F. Monte Carlo theory and practice. *Reports on Progress in Physics* **43** (1980).
56 Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**, 618-630 (2013).
57 Goncalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**, 835-849 e838 (2022).
58 Qiao, C. *et al.* Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nat Methods* **18**, 194-202 (2021).
59 Qiao, C. *et al.* Rationalized deep learning super-resolution microscopy for sustained live imaging of rapid subcellular processes. *Nat Biotechnol* **41**, 367-377 (2023).
60 Liu, Z. *et al.* Bioorthogonal photocatalytic proximity labeling in primary living samples. *Nat Commun* **15**, 2712 (2024).
61 Zhang, Z. *et al.* Progress, Challenges and Opportunities of NMR and XL-MS for Cellular Structural Biology. *JACS Au* **4**, 369-383 (2024).
62 Scarmeas, N., Anastasiou, C. A. & Yannakoulia, M. Nutrition and prevention of cognitive impairment. *Lancet*

Neurol **17**, 1006-1015 (2018).
63 Kottek, M., Grieser, J., Beck, C., Rudolf, B. & Rubel, F. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift* **15**, 259-263 (2006).
64 Harel, M. *et al.* Proteomics of Melanoma Response to Immunotherapy Reveals Mitochondrial Dependence. *Cell* **179**, 236-250 e218 (2019).
65 Xu, J. Y. *et al.* Integrative Proteomic Characterization of Human Lung Adenocarcinoma. *Cell* **182**, 245-261 e217 (2020).
66 Shi, Y. *et al.* Targeting LIF-mediated paracrine interaction for pancreatic cancer therapy and monitoring. *Nature* **569**, 131-135 (2019).
67 Eckert, M. A. *et al.* Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts. *Nature* **569**, 723-728 (2019).
68 Shen, B. *et al.* Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* **182**, 59-72 e15 (2020).
69 Nie, X. *et al.* Multi-organ proteomic landscape of COVID-19 autopsies. *Cell* **184**, 775-791 e714 (2021).
70 Niu, L. *et al.* Noninvasive proteomic biomarkers for alcohol-related liver disease. *Nat Med* **28**, 1277-1287 (2022).
71 Virreira Winter, S. *et al.* Urinary proteome profiling for stratifying patients with familial Parkinson's disease. *EMBO Mol Med* **13**, e13257 (2021).
72 Wigger, L. *et al.* Multi-omics profiling of living human pancreatic islet donors reveals heterogeneous beta cell trajectories towards type 2 diabetes. *Nat Metab* **3**, 1017-1031 (2021).
73 Johnson, E. C. B. *et al.* Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. *Nat Neurosci* **25**, 213-225 (2022).
74 Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat Commun* **9**, 882 (2018).
75 Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *Nat Protoc* **16**, 5398-5425 (2021).
76 Su, P. *et al.* Single Cell Analysis of Proteoforms. *J Proteome Res* **24**, 1883-1893(2024).
77 Chen, W. *et al.* Simple and Integrated Spintip-Based Technology Applied for Deep Proteome Profiling. *Anal Chem* **88**, 4864-4871 (2016).
78 MacCoss, M. J. *et al.* Sampling the proteome by emerging single-molecule and mass spectrometry methods. *Nat Methods* **20**, 339-346 (2023).
79 Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* **40**, 1231-1240 (2022).
80 Geyer, P. E. *et al.* Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst* **2**, 185-195 (2016).
81 Deutsch, E. W. *et al.* Advances and Utility of the Human Plasma Proteome. *J Proteome Res* **20**, 5241-5263 (2021).
82 Buljan, M. *et al.* A computational framework for the inference of protein complex remodeling from whole-proteome measurements. *Nat Methods* **20**, 1523-1529 (2023).
83 Mackmull, M. T. *et al.* Global, in situ analysis of the structural proteome in individuals with Parkinson's disease to identify a new class of biomarker. *Nat Struct Mol Biol* **29**, 978-989 (2022).
84 Tsamardinos, I. *et al.* Just Add Data: automated predictive modeling for knowledge discovery and feature selection. *NPJ Precis Oncol* **6**, 38 (2022).
85 Bai, Y. *et al.* AutoDC: an automatic machine learning framework for disease classification. *Bioinformatics* **38**, 3415-3421 (2022).
86 Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348-352 (2021).

---

[1]State Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China. [2]International Academy of Phronesis Medicine (Guangdong), Guangdong, 510320, China. [3]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland. [4]Macquarie Medical School, Macquarie University, North Ryde, Sydney, NSW, 2109, Australia. [5]Institute of Pathology and Southwest Cancer Center, Southwest Hospital, Third Military Medical University (Army Medical University); Key Laboratory of Tumor

Immunopathology, Ministry of Education of China, Chongqing, China. [6]Institute of Health Service and Transfusion Medicine, Beijing 100850, China. [7]Department of Pathology, Johns Hopkins University; The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland, USA. [8]Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China. [9]Department of Nephrology, First Medical Center of Chinese PLA General Hospital, Nephrology Institute of the Chinese People's Liberation Army, State Key Laboratory of Kidney Diseases, National Clinical Research Center for Kidney Diseases, Beijing Key Laboratory of Kidney Disease Research, Beijing 100853, China. [10]Institute of Chemistry, Academia Sinica, Taipei, China. [11]National Biomedical Imaging Center, State Key Laboratory of Membrane Biology, Institute of Molecular Medicine, Peking-Tsinghua Center for Life Sciences, College of Future Technology, Peking University, Beijing, China. [12]School of Biological Sciences, Queen's University of Belfast, Belfast, U.K. [13]Functional Proteomics Laboratory, Centro Nacional de Biotecnología-CSIC, Darwin 3, 28049 Madrid, Spain. [14]Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Martinsried, Germany. [15]AI for Science Institute, Beijing 100083, China, Center for Machine Learning Research, Peking University, Beijing 100871, China. [16]Advanced Clinical Biosystems Research Institute, Cedars Sinai Medical Center, Los Angeles, California, United States. [17]Department of Liver Surgery and Transplantation, Key Laboratory of Carcinogenesis and Cancer Invasion (Ministry of Education), Liver Cancer Institute, Zhongshan Hospital, Fudan University, Shanghai, China. [18]Department of Medicine, School of Clinical Sciences, Monash University, Clayton, Victoria, Australia. [19]School of Pharmaceutical Sciences and Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada. [20]The D. H. Chen School of Universal Health, Zhejiang University, Hangzhou, The People's Republic of China. [21]Pengcheng Laboratory, Shenzhen 518052, China, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China. [22]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada. [23]Department of Chemistry, University of Tokyo, Tokyo, Japan. [24]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore. [25]School of Medicine, Southern University of Science and Technology, Shenzhen, 518055, China. [26]Department of Nutrition, Tianjin Institute of Environmental and Operational Medicine, Tianjin, China. [27]School of Medicine, Westlake University, Hangzhou, Zhejiang Province, China; Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang Province, China; Research Center for Industries of the Future, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province, China. [28]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Padualaan 8, Utrecht, 3584 CH, the Netherlands. Netherlands Proteomics Center, Padualaan 8, Utrecht, 3584 CH, the Netherlands. [29]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [30]Molecular and Cell Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, United States. [31]Department of Head & Neck Surgery, Division of Surgery & Surgical Oncology, Division of Medical Sciences, National Cancer Centre Singapore, Singapore. [32]OncoProteomics Laboratory, Department of Medical Oncology, Cancer Center Amsterdam, Amsterdam UMC, Location Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV, Amsterdam, the Netherlands. [33]Advanced Glycoscience Research Cluster, School of Biological and Chemical Sciences, University of Galway, Ireland. [34]Departments of Molecular Biosciences, Departments of Chemistry, Northwestern University, United States. [35]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada. Central China Institute of Artificial Intelligence, Henan, China. [36]Department of Biological Sciences, Faculty of Science, National University of Singapore, Singapore. [37]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [38]Department of Structural Biology, Leibniz-Forschungsinstitut für MolekularePharmakologie (FMP), Robert-Roessle-Str. 10, Berlin 13125, Germany. [39]State Key Laboratory of Membrane Biology, Key Laboratory of Organ Regeneration and Reconstruction, Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101, China. [40]Cancer Biology Institute, Yale University School of Medicine, West Haven, CT, 06516, United States. [41]State Key Laboratory of Molecular Oncology, National Cancer Center, National Clinical Research Center for Cancer, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. [42]UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000 Kuala Lumpur, Malaysia. [43]Department of Critical

Care Medicine and Hematology, The 3rd Xiangya Hospital, Central South University; Department of Hematology and Critical Care Medicine, The second Xiangya Hospital, Central South University, Changsha, China. [44]Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany. [45]School of Life Sciences, Tsinghua University, Tsinghua-Peking Center for Life Sciences, Beijing 100084, China. [46]Institute for Systems Biology, Seattle, Washington 98109, United States. [47]Clinical Biomarker Discovery and Validation, Monash University, Clayton, Victoria, Australia. [48]Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; Shanghai National Clinical Research Center for Metabolic Diseases, Key Laboratory for Endocrine and Metabolic Diseases of the National Health Commission of the PR China, Shanghai Key Laboratory for Endocrine Tumor, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China [49]Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, United States. [50]Department of Oral Biological and Medical Sciences, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, Yonsei Frontier Lab, Yonsei University, Seoul 03722, Republic of Korea. [51]Glycoproteomics Laboratory, Department of Parasitology, University of São Paulo, Sao Paulo 05508-000, Brazil. [52]Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; Beijing Institute for Stem Cell and Regenerative Medicine, Beijing 100101, China; University of the Chinese Academy of Sciences, Beijing 100049, China. [53]Univ Rennes, Inserm, EHESP, Irset (Institut de Recherche en Santé, Environnement et Travail) - UMR_S 1085, Rennes, F-35042, France. [54]Pilot Laboratory, MOE Frontier Science Centre for Precision Oncology, Centre for Precision Medicine Research and Training, Institute of Translational Medicine, Faculty of Health Sciences, University of Macau, Macau 999078, China. [55]Infection and Immunity Program and Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Clayton, Victoria, Australia. [56]State Key Laboratory of Female Fertility Promotion, Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing 100191, China. [57]ProCan®, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia. [58]Department of Health Sciences, University Magna Græcia of Catanzaro, 88100 Catanzaro, Italy. [59]Department of Systems Biology, Harvard Medical School, Boston, MA, USA; Broad Institute of MIT and Harvard, Cambridge, MA, USA. [60]State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, China. [61]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China; Breast Tumor Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China. [62]Indian Institute of Technology Bombay, Mumbai, 400076, India. [63]Department of Health Sciences, Faculty of Applied Health Sciences, Brock University, St. Catharines, ON, Canada. [64]Center for Quantitative Biology, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. [65]Department of Chemistry, Southern University of Science and Technology, Shenzhen, China. [66]Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [67]Department of Neurology, Clinical Neuroscience Center, University Hospital Zurich and University of Zurich, Zurich, Switzerland. [68]Technical University of Munich, 85354 Freising, Germany. [69]National Laboratory of Genomics for Biodiversity UGA-Langebio, Center for Research and Advanced Studies (CINVESTAV) Irapuato, Km. 9.6 Libramiento Norte Carr. Irapuato-León, 36824, Irapuato, Gto, Mexico. [70]Institute of Translational Medicine, Department of Health Sciences and Technology, ETH Zurich, 8093 Zurich, Switzerland. [71]Department of Computer Science, National University of Singapore, 117417, Singapore; Department of Pathology, National University of Singapore, 119074, Singapore. [72]Guangzhou National Laboratory, Guangzhou, China. [73]School of Biomedical Engineering, Guangzhou Medical University, Guangzhou, China. [74]The Scripps Research Institute, La Jolla, CA, United States. [75]China Science and Technology Exchange Center, Beijing 100045, China. [76]CAS Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [77]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA; Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [78]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, United States. [79]State Key Laboratory of Medical Proteomics, National Chromatography R. & A. Center, CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China. [80]School of Mathematical Sciences, Peking University, Beijing, China; Wuhan

University, Wuhan, China. [81]A list of members and their affiliations appears at the end of the paper.
[*]Correspondence should be addressed to F.H. and R.A. (hefc@bmi.ac.cn; aebersold@imsb.biol.ethz.ch)

The π-HuB Consortium

Ruedi Aebersold[3], Mark S. Baker[4], Xiuwu Bian[5], Xiaochen Bo[6], Daniel Chan[7], Cheng Chang[1], Luonan Chen[8], Chris Soon Heng Tan[65], Xiangmei Chen[9], Yu-Ju Chen[10], Heping Cheng[11], Ben C Collins[12], Fernando Corrales[13], Jürgen Cox[14], Weinan E[15], Jennifer Van Eyk[16], Jia Fan[17], Pouya Faridi[18], Daniel Figeys[19], George Fu Gao[20], Mingxia Gao[82], Qiang Gao[17], Wen Gao[21], Yushun Gao[66], Zu-Hua Gao[22], Keisuke Goda[23], Wilson Wen Bin Goh[24], Dongfeng Gu[25], Changjiang Guo[26], Tiannan Guo[27], Xuejiang Guo[60], Fuchu He[1,2], Jie He[66], Jun He[83], Qing-Yu He[84], Yuezhong He[2], Albert J.R. Heck[28], Henning Hermjakob[29], Jinlin Hou[85], Canhua Huang[86], Tony Hunter[30], Chenxi Jia[1], Narayanan Gopalakrishna Iyer[31], Ying Jiang[1], Connie R Jimenez[32], Lokesh Joshi[33], Neil L Kelleher[34], Bernhard Kuster[68], Chaoying Li[1], Dong Li[1], Ming Li[35], Yan Li[87], Yanchang Li[1], Yang Li[1], Qingsong Lin[36], Cui Hua Liu[37], Fan Liu[38], Guang-Hui Liu[39], Siqi Liu[88], Xiaonan Liu[89], Ya Liu[89], Yansheng Liu[40], Zhihua Liu[41], Zhongyang Liu[1], Teck Yew Low[42], Ben Lu[43], Haojie Lu[82], Matthias Mann[44], Anming Meng[45], Robert L. Moritz[46], Edouard Nice[47], Yongzhan Nie[89], Guang Ning[48], Mariko Okada[90], Gilbert S Omenn[49], Christopher M Overall[50], Giuseppe Palmisano[51], Yaojin Peng[52], Charles Pineau[53], Terence Chuen Wai Poon[54], Anthony Purcell[55], Guojun Qian[91], Jie Qiao[56], Hongqiang Qin[79], Yu Rao[92], Zihe Rao[93], Roger Reddel[57], Xianwen Ren[94], Yan Ren[88, 95], Phil Robinson[57], Paola Roncada[58], Chris Sander[59], Jiahao Sha[60], Feng Shen[91], Lin Shen[87], Erwei Song[61], Sanjeeva Srivastava[62], Shicheng Su[61], Aihua Sun[1], Siu Kwan Sze[63], Minjia Tan[96], Ben Zhong Tang[97], Chao Tang[64], Liujun Tang[1], Sheng-Ce Tao[98], Ruijun Tian[65], Juan Antonio Vizcaíno[29], Chanjuan Wang[1], Chen Wang[66], Jian Wang[1], Tong Wang[84], Xiaowen Wang[1], Xinxing Wang[26], Yan Wang[1], Liming Wei[82], Tobias Weiss[67], Mathias Wilhelm[68], Robert Winkler[69], Bernd Wollscheid[70], Catherine C L Wong[66], Limsoon Wong[71], Linhai Xie[1], Wei Xie[45], Xiaoliang Sunney Xie[94, 99], Li Xu[100], Ping Xu[1], Tao Xu[73], Tianhao Xu[2], Liying Yan[56], Huanming Yang[88], Jianjun Yang[89], Jing Yang[72], Xiao Yang[1], John Yates[74], Mingliang Ye[79], Wantao Ying[1], Xiaobo Yu[1], Yaxiang Yuan[101], Qingcun Zeng[102], Qiwei Zhai[76], Qimin Zhan[87, 103], Bing Zhang[77], Hui Zhang[78], Lihua Zhang[79], Lingqiang Zhang[1], Pingwen Zhang[80], Xiaofei Zhang[83], Xu Zhang[104], Ying Zhang[82], Yukui Zhang[79], Yu Zi Zheng[22], Nan-Shan Zhong[72, 105], Qing Zhong[57], Feng Zhou[82], Yi Zhu[27], Yunping Zhu[1], Zemin Zhang[94, 99].

[82]Institutes of Biomedical Sciences, Fudan University, Shanghai, 200433, China. [83]Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China. [84]College of Life Science and Technology, Jinan University, Guangzhou, China. [85]Department of Infectious Diseases, Nanfang Hospital, Southern Medical University, Guangzhou, China. [86]State Key Laboratory of Biotherapy, West China Hospital, and West China School of Basic Sciences & Forensic Medicine, Sichuan University, Chengdu, China. [87]Peking University Cancer Hospital & Institute, Beijing, China. [88]BGI Group, Shenzhen 518083, P. R. China. [89]Xijing Hospital, Fourth Military Medical University, Xi'an, China. [90]Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan. [91]Eastern Hepatobiliary Surgery Hospital, Second Military Medical University (Naval Medical University), Shanghai, China. [92]School of Pharmaceutical Sciences, Tsinghua University, Beijing, China. [93]School of Medicine, Tsinghua University, Beijing, China. [94]Changping Laboratory, Beijing, China. [95]Experiment Center for Science and Technology, Shanghai University of Traditional Chinese Medicine, Shanghai, China. [96]State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. [97]School of Science and Engineering, Shenzhen Institute of Aggregate Science and Technology, The Chinese University of Hong Kong, Shenzhen, China. [98]Shanghai Center for Systems Biomedicine, Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Jiao Tong University, Shanghai, China. [99]Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing, China. [100]Department of Liver Surgery, Sun Yat-sen University Cancer Center, Collaborative Innovation Center for Cancer Medicine, Guangzhou, China. [101]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. [102]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. [103]Peking University International Cancer Institute, Beijing, China. [104]Department of Urology, The Third Medical Center, Chinese PLA General Hospital, Beijing, China. [105]State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China