National Advisory Council for Human Genome Research
May 18-19, 2020
Concept Clearance for RFA

Technology Development for Single-Molecule Protein Sequencing

Purpose:

The purpose of this initiative is to accelerate innovation, development and early dissemination of single-molecule protein sequencing (SMPS) technologies. The ultimate goal is to achieve technological advances to the level where protein sequencing data can be generated at sufficient scale, speed, cost and accuracy to use routinely in studies of genome biology and function, and in biomedical research in general. This would enable analyses, such as deeper understanding of molecular phenotypes, identification of low abundance and 'missing' proteins, and true single-cell proteomics. This concept represents an ambitious and high-risk technology development challenge, and if successful, would provide a focused opportunity to transform the use of proteomics, in much the same way as modern next-generation nucleic acid sequencing (NGS) transformed genomics due to its high-throughput, low cost, and generalizability.

Background:

Through the Human Genome Project, and other efforts, NHGRI transformed biology by making genomics mainstream, with genomics now being a fundamental part of many studies of the biology of disease. Proteomics has not had the same widespread adoption, partially because of a lack of proteomics technologies that approach the scale of NGS, including improvements to the sensitivity and dynamic range of protein detection.

The human proteome is extremely complex. A typical human cell expresses >10,000 unique protein gene products; and can contain ~100 times as many modified proteins, or proteoforms, for each gene product. In addition, the dynamic range of the proteome approaches seven orders of magnitude (from one copy per cell to ten million copies per cell) in tissues and cell lines, and up to ten orders of magnitude in blood plasma. Currently, the main approaches used to measure proteins are affinity reagent-based approaches and mass spectrometry (MS)-based approaches. While both are valuable, they have their limitations. Affinity approaches have excellent spatial resolution, can detect small amounts of protein, and can be used in intact tissues. Although they can be multiplexed, they rely on being aware of protein targets against which to develop what is essentially a set of custom reagents, thus they are limited in scale. MS is the mainstay technology for large-scale protein sequencing and identification, but there have been no revolutionary advances in MS for some time. MS technology lacks the sensitivity and dynamic range needed to routinely detect low-abundance proteins in human samples, and is limited in its ability to map proteoforms. The inability to comprehensively identify and quantify proteins in such complex human samples is a major roadblock in protein biomarker discovery and quantitative systems biology.

There is now a third emerging approach, single-molecule protein sequencing, based on technologies such as nanopore sequencing and Edman-like massively parallel peptide sequencing. These approaches have the potential to detect and quantify very small amounts of protein and proteoforms, and to approach the scale of NGS which would enable the analysis of complex human protein samples. SMPS is a far less developed approach compared to affinity and MS approaches; as such it presents a significant opportunity to advance the state-of-the-art. The realization of SMPS technologies would represent a disruptive burst in proteomics research by facilitating the detection of low abundance proteins and enabling true single-cell protein analysis at high throughput. Single-molecule detectors could also advance the search for the so-called "missing proteins". The human

proteome consists of ~ 3,000 proteins that have never been directly identified despite genetic or transcriptional evidence of existence. Thus, this technology could offer significant improvement to cataloguing protein gene products encoded in the human genome.

Several research priorities are emerging as part of NHGRI's strategic planning for a "2020 Vision for Genomics" which this initiative, if successful, is poised to advance. *First*, data types and molecular phenotypes beyond primary DNA sequence (such as readouts from RNA, proteins, metabolites) are needed to fully understand genome function in different biological contexts. *Second*, successful development of single-molecule protein sequencing would add an important data type to the molecular phenotyping toolbox for genomic analysis at single cell resolution. *Third*, high throughput detection of protein and proteoforms may help establish roles of all protein-coding genes in pathways and networks. *Lastly*, this technology may enable multi-omic approaches for the diagnosis and management of human disease. The high-sensitivity nature of the SMPS method could prove to be ground-breaking. Integrating such proteomic data with other -omic data, along with clinical variables and outcomes, should advance understanding of disease onset and progression, and could lead to improved predictive and prognostic models for a wide range of conditions.

Proposed Scope and Objectives:

NHGRI has a unique and long-standing commitment to fostering genomic technology development. NHGRI's investment in catalyzing the development of new DNA sequencing technologies was a major part of the underlying successes in reaching the '$1000 genome.' Over the years, NHGRI has broadened the scope of technology development beyond nucleic acid sequencing to include other areas such as technologies to determine genomic function, gene regulation, chromatin state, nuclear organization, and dynamics of those features in genomes of single cells, or mixed populations of cells. A natural extension would be to broaden beyond nucleic acid sequencing to amino acid sequencing of proteins.

Development of single-molecule technologies for protein sequencing is in its infancy and is extremely challenging; yet recent advances show high promise and thus, the field is ripe for investment. To our knowledge, there has been no funding opportunity available to date to advance this promising field. Although much of proteomics is being funded by other NIH institutes and centers, SMPS represents a focused contribution to the field of proteomics that is within the scope of NHGRI's mission, particularly in relation to ideas that have emerged from the NHGRI's strategic planning for a "2020 Vision for Genomics".

This program will support investigator-initiated research with an aim to significantly advance technologies for protein sequencing. All funded PIs will be expected to attend an annual meeting hosted by NHGRI to present and openly share their findings, while seeking opportunities to network and collaborate and learn from one another.

The following are some examples of the types of techniques for single-molecule protein sequencing that would be appropriate for development for these FOAs:

- nanopore sequencing
- Edman-like degradation with parallel measurements
- fluorescence-based techniques
- tunneling currents

Techniques that would not be considered appropriate for these FOAs.

- mass spectrometry
- affinity reagents
- techniques that do not have the potential for proteome-scale analysis

Relationship to Ongoing Activities:

The proposed initiative will uniquely address technology development for single-molecule protein sequencing. The initiative is intended to nurture and expand protein sequencing research by also enhancing interactions among grantees and promoting sharing of successful approaches and resulting data. Related applications might also be received through the NIH Parent R01 and R21 announcements, as well as Novel Genomic Technology Development PARs (PAR-18-777, PAR-18-778, PAR-18-779). Although these FOAs might receive some applications with relevance to this concept, none specifically call for protein sequencing projects.

There are only a handful of active grants focused on SMPS technology development being supported across the NIH. These investigators would be invited to participate in an annual meeting hosted by NHGRI to share their research and network within a community of technology development-minded investigators. Commercial interests are in the early stages of developing and deploying technologies in this area. To the extent possible, NHGRI will attempt to avoid overlap with those efforts, and leverage opportunities. Support for commercial efforts in these areas will be encouraged via the Small Business Program of NHGRI.

Mechanisms of Support (RFA):

- R01 (Research Project); up to $500K direct costs/year; project period of up to 3 years
- R21 (Exploratory/Developmental Research); up to $200K direct costs/year; project period of up to 2 years
- R43/R44 SBIR; up to $250K for Phase I, up to $2M for Phase II

Funds Anticipated:

Anticipated duration of the program is 5 years with an investment of $29M in total costs for this first round of funding. Per the table below, the program would ramp to $9M/yr. If, after a mid-course evaluation in year 3, it is decided that the program should continue, then funds will be maintained at $9M/yr. NHGRI is also seeking partnership with other ICs for this concept; as this technology would not only benefit the field of genomics, but would enable significant advances in both biomedical research and clinical settings.

	FY21	**FY22**	**FY23**	**FY24**	**FY25**
R01	2	4	6	4	2
R21	0.5	1	1	0.5	
SBIR	1	2	2	2	1
Total	**3.5**	**7**	**9**	**6.5**	**3**
Total Cost in millions					Grand Total = $29M