What is the mission of the Chromosome-centric Human Proteome Project (C-HPP?)
We have started with an easy question to answer. The Human Proteome Project (HPP) of the Human Proteome Organization (HUPO) aims to find high-stringency evidence for all proteins encoded by the human genome, the major splice forms of each protein, mature N- and C-termini, and their major protein post-translational modifications (PTMs). Indeed, one cannot do the best biology and/or medical research without a complete understanding of the parts of the human proteome. Conversely, the international C-HPP teams need the input of biology/disease (B/D) teams to understand the biological context of the parts list. It is a question of focus. A focus on the parts list promotes a measurement- and detail-focused analytical mindset, while a focus on the biological context is focused on outcomes and may ignore individual parts in the context of the overall picture. In conclusion, both mindsets are indivisible parts of the larger HPP.

What is a Missing Protein?
Amazingly, as of 11th January 2019, 2,129 proteins representing 10.7% of all 19,823 human proteins recognized by neXtProt, are currently defined as 'missing' without compelling mass spectrometry (MS) or other evidence of protein existence (PE). These proteins have classification as PE2 – 4. Moreover, a surprising 1,057 PE1 proteins have only been identified by protein chemistry or other techniques, and still lack convincing evidence at the MS level (meeting the HPP Guidelines 3.0) with ≥ 2 proteotypic, non-nested peptides, ≥ 9 amino acids in length. Such PE1 entries that are lack any or complete MS evidence are non-MS PE1 proteins (proposed nomenclature).

Why do some proteins elude detection by mass spectrometry?
Various factors influence the ability of MS-based technologies to unambiguously detect missing and non-MS PE1 proteins. Conventional bottom-up workflows utilize trypsin to digest samples prior to LC-MS/MS analysis. Thus, a typical proteomics experiment is unlikely to detect proteins that are not amenable to trypsin digestion, or proteins which yield tryptic peptides that are difficult to ionize, fall outside of the range of typical mass detection ranges, or are not proteotypic (i.e. unique to a particular protein). Furthermore, proteins that are extensively crosslinked, insoluble or expressed in very low amounts or in a very limited temporal-spatial manner or found in challenging tissues to extract for analysis, for example mineralized tissues (bone, dental enamel, dentine and cementum), also present challenges for detection by MS. Other factors are limitations in mass spectrometry instrumentation and search algorithms—despite overlap, various search engines will preferentially identify different subsets of peptides.

Are missing proteins missing because they are unimportant?
No. The olfactory receptors are a compelling example of the lack of our knowledge in important areas of biology. These receptors represent a large class of proteins whose coding regions appear to be distributed over almost all the chromosomes. The proteins are extensively expressed in nasal tissue but there is also unconfirmed evidence at the protein level that olfactory receptors may also be expressed in other tissues, ranging from reproductive organs to tooth dentine to the brain, and they appear to have important interactions with the immune system. Another huge example are the hundreds of ORF genes for which tissue expression of mRNA (PE2) has been demonstrated. MPs are also expressed in less well studied cells and tissues that may provide phenotypic differences to these cell populations and tissues, but are difficult to sample or access.

Have we finished the C-HPP once we have found all the missing proteins?
No. Finding all MPS is just Phase I of the C-HPP. Next comes functionalization of the PE1 proteins with no known or predicted function, or homology or orthologues with known function. Extensive goals such as ASV and PTM characterizations follow; thus, each team has many proteomic adventures ahead. Our objective across the HPP will be to identify and characterize the dynamics, functions, associations and turnover of their numerous isoforms and variants, their evolution, and their roles in pathways, networks, embryogenesis, growth and development, and disease processes.

How is proteomic data captured and used for the C-HPP?
While a quality review is the foundation of a scientific study, the acceptance of a proteomic study for publication, even in high impact journals, does not guarantee the use of its dataset. Many experimental details and lab-to-lab variations make very difficult the comparison of the datasets from different research groups, especially in a global collaboration like the C-HPP. For the C-HPP to succeed, the data must be curated and accepted at either PeptideAtlas (http://www.peptideatlas.org) and/or now at MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) and integrated into the databases. Furthermore, our gold standard for the C-HPP is acceptance with gold-level evidence in neXtProt (https://www.nextprot.org).

Once a dataset has been deposited in ProteomeXchange, is the task complete?
No. The authors must also provide the PXD identifiers in the publication, preferably also in the abstract, and then provide permission for the dataset to be made public. To facilitate the dataset's use by other teams, additional steps should be considered to guide each team to access the C-HPP Wiki (https://c-hpp.web.rug.nl/tiki-index.php?page=HomePage) and distribute the produced dataset with biological and other details. MissingProteinPedia (http://www.missingproteins.org/) also contains orthogonal MP evidence.

How is a Missing Protein identified for the HPP?
Peptides and proteins are identified in proteomic data sets, but only those reaching very high stringency in spectral quality and sequence coverage are considered as unequivocal evidence for the promotion of a PE2-4 protein (missing protein) to a PE1 protein (i.e. found). Full details of the criteria used to designate a missing protein as found are those meeting the current HPP Guidelines (v 3.0 in press) (https://hupo.org/Guidelines, and for v2.1 DOI: 10.1021/acs.jproteome.6b00392, J. Proteome Res. 2016, 15, 3961−3970).

Notably, while all peptides that pass thresholds are visible in PeptideAtlas and neXtProt, only the proteins with two uniquely-mapping non-nested proteotypic peptides with length 9AA or greater, as called by neXtProt, are deemed to have sufficient evidence to be labeled as confidently detected by MS methods. Therefore, neXtProt is the final arbiter to designate if a PE2 – 4 protein in UniProtKB is now deemed PE1 in neXtProt for the purposes of the HPP.

What are the details of the HPP Data Processing Pipeline?
The current basic process by which the HPP manages the process of reducing the number of missing proteins of the human proteome begins with the collection of mass spectrometry data sets and deposition in one of the ProteomeXchange repositories after registration with ProteomeXchange. From their ProteomeXchange deposition, PeptideAtlas collects the raw data files and reprocesses those data using the tools of the Trans-Proteomic Pipeline (TPP). In 2019 the C-HPP is now also utilizing the MassIVE database. Thresholds are set extremely high in PeptideAtlas in order to obtain a 1% protein-level FDR across the ensemble of all datasets. In November each year, PeptideAtlas stops processing new datasets and creates a build reflecting

the current state of the human proteome with MS evidence. In December the final peptide list is transferred to neXtProt for integration into its next build, usually in January or February each year. The official numbers to determine progress on the HPP are taken from the January/February neXtprot release, even though neXtprot is updated several times a year, which also change the numbers of proteins in each of the PE categories.

Does the C-HPP study biology and/or disease?
Virtually all proteomic experiments are performed in a chromosome agnostic manner. Further, most of the chromosome teams conduct at least some of their search for the missing proteins (PE2 – 4) in samples relevant to biology or disease that allows the observation of unusual protein expression patterns that will provide unique insights for other biological studies. Many teams employ clinical studies to search for 'rare' proteins that are observed only in the unusual context of a disease state or rare human cells and tissues. Examples of teams supplying samples and data between groups include the US and ANZ teams on chromosome 17 and 7 and the Chinese teams (Chr 1, 8, 20) and Korean teams (Chr 9, 11, 13).

Does each chromosome team study only one chromosome?
Because proteomics experiments study proteins expressed potentially form every gene, the data generated are relevant to all the chromosome teams and thus must be shared with all the groups. Furthermore, a team's B/D studies will involve all the chromosomes, and these insights will be shared across teams. A focus of the C-HPP teams can also be on the parts list and biology of co-expression or *cis*-regulation of genes on that particular chromosome. Such biological features, exemplified by the HER2/neu (ERBB2) amplicon at chromosome 17q12, reinforces the scientific rationale for chromosome-by-chromosome studies of the proteome.

Is the value of a team defined by the number of missing proteins that they discover?
While the number of proteins discovered is an important metric, the data are only valuable once deposited and accepted by Peptide Atlas, MassIVE, and neXtProt. Now that C-HPP is well established, there are new opportunities to collaborate with the B/D groups to begin to understand the function of previously missing parts of the proteome. This is particularly relevant for functionalization of proteins, both MPs and PE1s, that have no known or predicted function, termed uPE1 proteins. Such flexibility will promote additional interactions between chromosome teams and B/D-HPP teams and allow for differences in priorities for various national and international funding agencies.

How does the C-HPP facilitate interactions and collaborations?
To facilitate productive and mutually beneficial interactions and collaborations within the C-HPP and between the C-HPP teams and those in B/D-HPP and the four support pillars of the HPP:

- Established an open structure where any group is free to work on any protein no matter where there are located in collaboration with the corresponding chromosome team. This principal was established at the Berlin Workshop, 2013.
- The C-HPP portal and the C-HPP wiki facilitate the translation and addition of information by researchers across the HPP in an open and transparent manner.
- The annual Special Issue of the Journal of Proteome Research dedicated to the C-HPP is now open to consider papers form all groups of the HPP.
- The semi-annual C-HPP workshops feature sessions and presentations from the B/D-HPP and the pillars.
- Most importantly, we are very collegial and friendly! Let's Go!

Christopher M. Overall (Co-Chair C-HPP), Bill Hancock (Past Chair C-HPP) and contributions modified from several relevant papers on which we are authors.