High throughput generation of a resource of the human secretome in mammalian cells

Tegel, Hanna; Dannemeyer, Melanie; Kanje, Sara; Sivertsson, Åsa; Berling, Anna; Svensson, Anne Sophie; Hober, Andreas; Enstedt, Henric; Volk, Anna Luisa; Lundqvist, Magnus

Total number of authors: 30

Published in:
New Biotechnology

Link to article, DOI:
10.1016/j.nbt.2020.05.002

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Tegel, H., Dannemeyer, M., Kanje, S., Sivertsson, Å., Berling, A., Svensson, A. S., Hober, A., Enstedt, H., Volk, A. L., Lundqvist, M., Moradi, M., Afshari, D., Ekblad, S., Xu, L. L., Westin, M., Bidad, F., Schiavone, L. H., Davies, R., Mayr, L. M., ... Hober, S. (2020). High throughput generation of a resource of the human secretome in mammalian cells. New Biotechnology, 58, 45-54. https://doi.org/10.1016/j.nbt.2020.05.002

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Full length Article

High throughput generation of a resource of the human secretome in mammalian cells

Hanna Tegel, Melanie Dannemeyer, Sara Kanje, Åsa Sivertsson, Anna Berling, Anna-Sophie Svensson, Andreas Hober, Henric Enstedt, Anna-Luisa Volk, Magnus Lundqvist, Mona Moradi, Delaram Afshari, Siri Ekblad, Lan Lan Xu, Malin Westin, Faranak Bidad, Lovisa Holmberg Schiavone, Rick Davies, Lorenz M. Mayr, Sinead Knight, Sven O. Göpel, Björn G. Voldborg, Fredrik Edfors, Björn Forström, Kalle von Feilitzen, Martin Zwahlen, Johan Rockberg, Jenny Ottosson Takanen, Mathias Uhlen, Sophia Hober, Sophia Hober*

*Department of Protein Science, School of Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden
bScience for Life Laboratory, KTH - Royal Institute of Technology, Stockholm, Sweden
cDiscovery Biology, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden
dDiscovery Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
eSyncona Investment Ltd, London, UK
fETH Zurich, B-BIOL, Molecular Health Sciences, Zurich, Switzerland
gBioscience Metabolism, Research and Early Development, Cardiovascular, Renal and Metabolism, BioPharmaceutical R&D, AstraZeneca, Gothenburg, Sweden
hNovo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Denmark

ARTICLE INFO

Keywords:
secreted proteins
protein production
protein purification
high-throughput

ABSTRACT

The proteins secreted by human tissues and blood cells, the secretome, are important both for the basic understanding of human biology and for identification of potential targets for future diagnosis and therapy. Here, a high-throughput mammalian cell factory is presented that was established to create a resource of recombinant full-length proteins covering the majority of those annotated as ‘secreted’ in humans. The full-length DNA sequences of each of the predicted secreted proteins were generated by gene synthesis, the constructs were transfected into Chinese hamster ovary (CHO) cells and the recombinant proteins were produced, purified and analyzed. Almost 1,300 proteins were successfully generated and proteins predicted to be secreted into the blood were produced with a success rate of 65%, while the success rates for the other categories of secreted proteins were somewhat lower giving an overall one-pass success rate of ca. 58%. The proteins were used to generate targeted proteomics assays and several of the proteins were shown to be active in a phenotypic assay involving pancreatic β-cell dedifferentiation. Many of the proteins that failed during production in CHO cells could be rescued in human embryonic kidney (HEK 293) cells suggesting that a cell factory of human origin can be an attractive alternative for production in mammalian cells. In conclusion, a high-throughput protein production and purification system has been successfully established to create a unique resource of the human secretome.

Introduction

A major impediment in biological sciences today is the accessibility of well-validated full-length proteins to explore their characteristics and functions. In order to increase understanding of the inherent control of the human body, knowledge of the human secreted proteins must be developed. Such proteins can be used in a number of different assays in order to explore their functions, analyze their structures and acquire knowledge of their interaction partners. The generation of proteins on a whole proteome scale requires the selection and development of unit
operations suitable for high-throughput production, enforcing the need to address strategic and technical issues, such as the choice of production host, expression system and downstream purification process; validation of the generated proteins also needs to be taken into account.

The human secretome comprises ca. 20,000 non-redundant proteins (www.ensembl.org), defined as one representative isoform from every gene locus. Of these 20,000, more than 2,500 belong to the group of proteins that are predicted to be secreted from the cell, the secretome. In this paper, the human secretome is defined according to the Human Protein Atlas classification [1], and includes genes coding for at least one protein isoform having a signal peptide and lacking a transmembrane region. In addition, several proteins that according to the UniProt database (www.uniprot.org) are annotated as secreted, despite lacking a signal peptide, are included. From these criteria, it has been estimated that the human secretome comprises 2,641 genes, 13% of all human genes [2,3]. A comprehensive annotation with the aim of providing information about the final localization of secretome proteins has revealed that of the 2,641 secreted proteins, 932 are secreted into intracellular vesicles or bound to the cell membrane. Hence, a large fraction of the proteins in the secretome are predicted not to be secreted extracellularly, but are instead retained intracellularly or within the plasma membrane [2]. Among the 1,709 proteins that are secreted extracellularly, many are predicted to remain in the close vicinity of the secreting cell, while 730 of these are predicted to be secreted into the blood. The latter can be divided into two groups depending on the route of secretion. They may either be secreted through the normal cellular secretion pathway, released by induced vesicular secretion, or cleaved from the cell surface through active release. In this paper, these proteins are divided into two groups: the first belong to the group designated “Blood” and the second to the group designated “Blood – other main location”, see definitions in Table 1. Another group of proteins also included in the “Blood – other main location” category consists of predicted secreted proteins expressed by blood cells, but also high expression in other tissues. To further build knowledge about the secreted human proteins, a resource of full-length proteins is instrumental.

Here, we report on the “Human Secretome Project” (HSP), a program in which synthetic constructs for each gene corresponding to a predicted secreted protein as well as 542 extracellular domains (ECDs) have been generated and used for protein expression in mammalian cell factories. The ECDs are included since these are secreted by the same machinery as the majority of the secretome. Furthermore, it is a very interesting group of proteins, both to understand cell biology but also for the development of pharmaceuticals. The expressed proteins are affinity purified in a high-throughput setting and thereafter analyzed regarding purity, yield and identity. Expression data in a mammalian recombinant host cell, CHO, for 2,189 gene constructs are reported which allow analysis of the relationship between protein characteristics and yield. Moreover, data are reported on an alternative expression system, in the human cell line HEK 293, for a selection of the clones that show degradation patterns when produced in CHO-cells. The proteins produced have been used for development of proteomic analysis assays based on experimental mass spectrometry (MS), including retention times and fragmentation spectra for the relevant peptides. Finally, the purified proteins have been used in a phenotypic assay to identify factors that affect dedifferentiation of β-cells. In conclusion, we provide a knowledge resource to facilitate basic and applied research covering the proteins actively secreted in human cells, tissues and organs.

### Materials and methods

For more detailed methods, please see the Supplementary information.

### Construct design

Constructs were designed based on sequence information found in the UniProt and Ensembl databases. For classical secreted proteins, the endogenous signal peptide was replaced with the CD33 signal peptide. For non-conventional secreted proteins, not having a predicted signal peptide, the CD33 signal peptide was added to the N-terminus after removal of the starting methionine in the sequence. For single-pass transmembrane proteins, the sequence of the extracellular domain was selected and the CD33 signal peptide was either added to, or used to replace, any endogenous signal peptide. All constructs were equipped with a Protein C purification tag at the C-terminus preceded by a TEV protease site [4], inserted to allow for cleavage of the tag. In all cases any predicted C-terminal propeptide was excluded from the construct and for proteins with a predicted N-terminal propeptide, constructs were designed both including and excluding the propeptide.

### Plasmid preparation

All designed constructs were synthesized and cloned into the expression vector pQMCF-1-MCS (Icosagen Cell Factory OÜ, Tartu, Estonia). The constructs were then expressed in the mammalian cell line HEK 293. The expression was induced by doxycycline, with a final concentration of 1 μg/mL.

### Table 1

Summary of the definitions of the different annotation categories

| Category                  | Definition                                                                 |
|---------------------------|---------------------------------------------------------------------------|
| Blood                     | Proteins secreted into the blood                                          |
| Blood - other main location| Proteins secreted into the blood with their main function related to another location, for example, membrane receptors with proteolytically cleaved secreted form or proteins secreted from blood cells having a majority of the expression in other tissues |
| Digestive system          | Proteins locally secreted to the gastrointestinal tract, for example, digestive enzymes |
| Male locally              | Proteins locally secreted in male tissues, for example, proteins related to sperm maturation |
| Female locally            | Proteins locally secreted in female tissues, for example, the oocyte zona pellucida proteins |
| Brain locally             | Proteins locally secreted in the brain, for example, some neuropeptides   |
| Matrix                    | Locally secreted to or in the extracellular matrix, including both structural and matricellular proteins |
| Other locally             | Proteins locally secreted by either specific tissues not included above, for example, skin, lung, eye and ear, or by a specific cell type, or by several different tissues |
| Secreted - no data        | Secreted proteins with little data and for which the location cannot be determined |
| ECD                       | The extracellular domain of single-pass transmembrane proteins             |
Estonia) by GeneArt (Thermo Fisher Scientific, Waltham, MA, USA). The plasmids were transformed into Top10 cells using standard procedures. Single colonies were chosen for cultivation and prepared using the Plasmid Plus midi Kit (Qiagen, Hilden, Germany) according to the manufacturer’s instructions.

Production of secreted proteins in CHO cells

All secreted proteins from CHO cells were produced by using a transient expression system from Icosagen Cell Factory; the QMCF Technology (Icosagen Cell Factory OÜ).

Small-scale

The cells were pelleted and resuspended in medium containing plasmid DNA and transfectected using a BTX electroporator (twin wave HT96 well system gemini X2, Harvard apparatus, Holliston, MA, USA). After transfection the cells were added to fresh pre-warmed medium containing penicillin-streptomycin and grown in fed batch for 13 d. 48 h after transfection the cells were diluted and after six days protein production was promoted by adding feed and shifting the temperature from 37 to 30 °C. A second feed was added 9 d after transfection. At day 13 the supernatant was clarified by centrifugation and stored at −20 °C.

Medium-scale

Cells were transfected by chemical transfection using reagent 007 (Icosagen Cell Factory OÜ). After 20-24 h, pre-warmed medium was added. Protein production was promoted after 72 h by adding 10% Basic Feed (Xell AG) and a temperature decrease to 30 °C. Feeding was continued at day 6, day 8 and day 10. At day 13 the supernatant was clarified by centrifugation and stored as described above.

Pilot-scale

Cells were transfected as described in the small-scale section. The transfectants were selected after at least 14 d by diluting the cells every second day with Geneticin (Gibco, Thermo Fisher Scientific) starting at 48 h. The culture volumes were simultaneously increased to the desired final production volume was reached. To increase protein production, the temperature was shifted to 30 °C while proceeding with production, the temperature was shifted to 30 °C while proceeding with production, the temperature was shifted to 30 °C while proceeding with production.

Production of secreted proteins in HEK 293 cells in small-scale

For transient production Expi293 F cells (Thermo Fisher Scientific) were used. Plasmid DNA was diluted in expression medium and added to 1 ml PEI MAX (Polysciences, Inc, Warrington, PA, USA) (1 mg/ml) and incubated for 15 min before addition to the cells. 24 h after transfection, cultures were diluted with expression medium. Harvesting was performed by centrifugation 4 d after transfection. The cultures were stored as described above.

Analysis of protein production

The CHO cell cultures were assessed for protein secretion (“first analysis”) 6 d post-transfection (3 d for medium-scale) using Western Blot (WB) analysis and at harvest using WB and sodium dodecyl sulfate – polyacrylamide gel electrophoresis (SDS-PAGE). The HEK cell cultures were analyzed at harvest using WB and SDS-PAGE. In WB a primary rabbit antibody against the C-tag, GTX18591 (GeneTex, Irvine, CA, USA) was used for protein detection.

Protein purification and analysis

Purification of the produced proteins was performed on ÄKTAexpress systems (GE Healthcare) with affinity chromatography using a Protein C-tag antibody matrix [5,6] packed in 1 ml HiTrap columns, followed by buffer exchange using 2 x 5 ml HiTrap desalting columns (GE Healthcare) at standard flow rates. Buffers used were wash buffer (20 mM Tris, 150 mM NaCl, 2 mM CaCl₂, pH 7.5) and elution buffer (20 mM Tris, 100 mM NaCl, 5 mM EDTA, pH 7.5). After filtration of the supernatant using a 0.45 μm syringe filter and addition of CaCl₂ to 2 mM, the samples were loaded onto the column and unbound proteins were washed out. Elution was performed with elution buffer and collected in the loops of the ÄKTAexpress system before buffer exchange to 1x PBS (2 mM NaH₂PO₄, 8 mM Na₂HPO₄, and 150 mM NaCl, pH 7.4). Proteins produced at medium-scale were purified similarly but with larger affinity columns (2 x 5 ml per 200 ml supernatant), a HiPrep 26/10 desalting column and adjusted flow rates.

Protein concentrations were determined using absorbance at 280 nm and analysis of protein purity and identity were performed by SDS-PAGE analysis and WB including deglycosylation of the proteins using Mix II (NEB, Ipswich, MA, USA) according to manufacturer’s instructions.

Production of standards for targeted proteomics analysis (QPrESTs and QTag)

The production of both HisABPOOneStrep (QTag) and stable isotope-labeled protein fragments (QPrESTs) used for protein quantification was essentially done as described in [7]. Absolute quantification of the QTag was obtained by amino acid analysis. The QPrESTs were produced in an auxotrophic E. coli strain according to [8]. The expressed QPrESTs were purified by the standard workflow used within the Human Protein Atlas for PrEST production [9].

Preparation of samples for protein identification using mass spectrometry (MS)

The secretome protein products were diluted to a final concentration of 1.5 μM in 50 mM ammonium bicarbonate into a 96-well plate. 15-30 pmol of each protein was transferred to a new 96-well plate and mixed with the corresponding isotope-labeled standard and quantification tag. The samples were reduced (5 mM DTT, 30 min at 56 °C), alkylated (10 mM 2-chloroacetamide/2-iodoacetamide, 30 min in the dark at RT) and cleaved with 200 ng of proteases grade porcine trypsin (Sigma-Aldrich) overnight at 37 °C and thereafter quenched by addition of formic acid (FA) to a final concentration of 1%. After digestion the samples were vacuum dried and stored at −20 °C.

Analysis using data-dependent acquisition mass spectrometry

The samples were analyzed using one of two different liquid chromatography (LC)-system setups. The first setup used a Dionex Ultimate 3000 (Thermo Fisher Scientific) equipped with a trap column and a 25 cm analytical C-18 column (Thermo Fisher Scientific). For the mobile phase, solvent A (3% acetonitrile (ACN), 97% H₂O, 0.1% FA) and solvent B (95% ACN, 5% H₂O, 0.1% FA) were used. Peptides corresponding to 1 pmol per protein were separated using a gradient of 4-40% solvent B over 9 min, 0.5 μl/min in solvent A. The second setup used a Dionex Ultimate 3000 equipped with a 15 cm analytical C18-column (Thermo Fisher Scientific). Peptides corresponding to 4 pmol per protein were separated using a gradient of 5-37% solvent B over 8 min, 150 μl/min. Both separation setups were connected to a Bruker Impact II (Bruker Daltonics, Billerica, MA, USA) and the samples were analyzed in data-dependent acquisition mode, with a 3 s cycle time. The method performed a survey scan from 150 to 2,200 m/z (1 Hz) followed by MS/MS scans acquired dynamically (2,500 cts =8 Hz to
25,000 cts = 32 Hz). The dynamic MS/MS acquisition selected ions with a charge state of 2-5 and implemented a smart exclusion (5x) set to 30 s.

**Data analysis of data-dependent acquisition results**

The raw data obtained were searched using MaxQuant [version 1.5.7.0] [10] to confirm the identity of the proteins. MS/MS spectra were searched in batches against a database containing all proteins produced in the same cultivation batch using Andromeda [11] with the entire CHO proteome as background (UP00001075, retrieved 2018.03.27, 23,888 entries) and a list of the most common contaminants. The multiplicity was set to 2, allowing for quantification against the stable isotope-labeled standard, with Arg10 and Lys8 selected as heavy labels. The false discovery rate (FDR) was set to 1% both at peptide and protein level and the minimum peptide length was set to 4.

**Data independent acquisition (DIA) library generation**

Peptides from the previously prepared MS samples were pooled in sets of eight proteins and analyzed using the same LC-MS/MS setup as for the analysis of protein products. Peptides were loaded onto a trap column, washed for 5 min with 100% of solvent A, separated on a 25 cm analytical C18 Easy-Spray column gradient of 6-28% solvent B and analyzed using a Top5 data-dependent acquisition (DDA) method. Raw files were searched in MaxQuant version 1.5.2.8 against the secretome protein product sequences and the MS/MS files from the MaxQuant searches were used to build a spectral library in Skyline [22]. All protein sequences can be found in the Panorama repository (https://panoramaweb.org/human_secretome, username: panorama + kthuhlen@proteinms.net).

**DIA analysis of pooled plasma samples**

Two different pools of human plasma samples obtained from healthy donors were digested as described in [7]. The analysis was performed on the same instrument setup and LC-gradient as described above for DIA library generation, but two of the samples were analyzed using a 50 cm C18 EASY-Spray column (Thermo Fisher Scientific) and the MS was operated in a DIA mode. For all samples a total of 1 μg of peptides was injected onto the column. One sample was also injected 12 times and analyzed in DIA using small isolation windows, and with the multiple injections covering the same range as the other DIA methods. Raw data files were imported into Skyline and matched against the curated spectral library. Peaks with matching fragmentation spectra were integrated and MS2 peak intensities were extracted for peptides that could be detected in the plasma samples. Extracted ion chromatograms were uploaded to Panorama, access as above.

**EndoC-βH1 Dedifferentiation screening assay**

EndoC-βH1 cells (Univercell Biosolutions, Toulouse, France) were cultured according to [12]. The cells were dispensed at a density of 6 × 10^3/well and incubated under standard culture conditions for 24 h. Cells were then treated with fibroblast growth factor FGF2 and neutrophil control (media) according to [13] or with the secretome library according to [14]. The secretome library applied consisted of 812 protein samples (corresponding to 765 unique genes) in 3-point concentration response. Dosed cells were incubated for another 96 h before fixation, permeabilization and labelling with MAF BZIP Transcription factor A (MAFA) (catalog no. #79737 Cell Signaling Technology, Inc., Danvers, MA, USA) and SRY-Box Transcription factor (SOX9) (catalog no. ab196184 Abcam PLC, Cambridge, UK) antibodies. Confirmation screening was carried out in a 10-point concentration response. All images were acquired using CV7000 (Yokogawa) confocal microscopy at 20X using BP filter 445/50 for Hoechst and BP filter 525/50 for MAFA and SOX9. Image segmentation and analysis were performed using Columbus™ software (PerkinElmer, Waltham, MA, USA).

**Single-Cell RNA-Seq of Pancreatic Islets**

In brief, human tissue and primary islets were purchased from Prodo Laboratories Inc. (Irvine, CA, USA). The use and storage of human islets and tissue samples were performed in compliance with the Declaration of Helsinki, ICH/Good Clinical Practice and was approved by the independent Regional Ethics Committee. Human islet samples (85%–95% pure) were cultured for 4 days in complete Prodo Islet Media Standard to recover after arrival. Islets were dissociated and distributed by Fluorescence Activated Cell Sorting into 384-well plates. Single-cell RNA-seq libraries were produced with the Smart-seq2 protocol [15]. Sequencing was carried out on an Illumina HiSeq 2000 generating 43 bp single-end reads. Sequence reads were aligned toward the human genome (hg19 assembly) using STAR (v2.3.0e), and uniquely aligned reads within RefSeq gene annotations were used to quantify gene expression as RPKMs using rpkmforgenes [16]. For a more detailed description see [15].

**Results and Discussion**

**A protein factory for production of the human secretome**

For production of the human secreted proteins and selected ECDs of single pass transmembrane proteins, a standardized protein production pipeline was set up (Fig. 1A). The system was based on a mammalian CHO cell host system in combination with semi-stable transfection of clones generated by gene synthesis. To enable production and subsequent purification of all proteins, the recombinant proteins were synthesized with an N-terminal signal peptide (CD33) and a C-terminal purification handle (Fig. 1B). After transfection of the expression vectors the proteins were transiently produced in mammalian CHO cells using the QMCF Technology [17] and subsequently purified using an antibody-based chromatography resin with a calcium ion-dependent affinity for the Protein C-tag, included at the C-terminus of the recombinant protein. This tag enables mild elution by the use of a cheleting elution buffer [18]. The purity and protein identity of the various target proteins were analyzed by SDS-PAGE, WB and MS/MS (Fig. 1A).

Initially, protein production was performed in a small-scale setting with a final culture volume of 50 ml (Fig. 1C). The recombinant proteins were purified from the conditioned medium with a protocol using the C-terminal purification tag. For the proteins produced at small scale, 1 ml of affinity matrix was used for each culture. The mean protein amount achieved in small scale was 755 μg for all proteins successfully produced (Fig. 2B). A maximum amount of 5 mg pure protein was achieved for C-X-C motif chemokine 5 (CXCL5) from a single 50 ml culture. However, depending on the intended application, larger protein amounts might also be needed. Therefore, a medium-scale protocol was developed (Fig. 1D), for which the final culture volume was 900 ml. This large volume of conditioned medium was purified using 40 ml of affinity matrix and normally generated over 10 mg of pure protein with a maximum, so far, of 153 mg. The mean protein amount for all proteins successfully produced at medium scale was 35 mg (Fig. 5A). Finally, to be able to produce even larger amounts of proteins, a protocol for a WAVE Bioreactor system with 20 L Cellbags and a final culture volume set to 10 L was developed (Fig. 1E). With this setup, a stable pool was generated prior to protein production and up to 1 g of purified protein could be produced.

**Small-scale production**

By using the small-scale protocol (Fig. 1C), 2,189 different genes have been initiated in the production pipeline using CHO cells as the
production host (Supplementary Table S1). From this high-throughput production system without optimization for individual proteins 1,276 different proteins (58%) were successfully produced and purified. The success rate for each annotated category, see definitions in Table 1, is shown in Fig. 2A and interestingly the highest success rate was achieved when producing proteins that are naturally secreted to the digestive system and to blood, 78% and 66%, respectively. Also, the selected extracellular domains (ECDs) showed a high success rate (71%). Proteins with the lowest success rate regarding production and purification are annotated as being secreted to the cell matrix (33%). Furthermore, the group of proteins that is less understood and comprises proteins that remain to be explored is also among those with very low success rate (37%).

The protein yields after production and purification varied among the proteins (Fig. 2B). Many of the secreted proteins are post-translationally modified during translocation through the secretory pathway. In order to assess the size and purity of the produced proteins and also to achieve information on the degree of glycosylation, a mixed deglycosylation enzyme kit was used. This is exemplified in Fig. 2C showing heterogeneous and oversized bands for two of the proteins after CHO production and single bands of expected size after deglycosylation. As expected, the CHO host system often yields proteins with differential glycosylation patterns, a phenomenon that has been reported previously [19].

Fig. 1. A summary of the different production pipelines is shown. A) An outline of the protein production pipeline including production, purification and analyses. The first step in the standardized high-throughput protein production pipeline used in the HSP is construct design. The constructs are then synthesized and cloned into the expression plasmid. All plasmids are prepared and sequence verified before protein production. The Protein C-tagged target proteins are then purified using an automated affinity purification setup. Purified proteins are identified and quantified with MS/MS and purity and glycosylation patterns are determined using SDS-PAGE and WB. B) All proteins produced in the HSP are produced with an N-terminal signal peptide, CD33, and a C-terminal purification handle based on the Protein C-tag purification tag. Between the purification tag and the protein, a TEV protease site was inserted to allow for cleavage of the tag, if needed. The CMV promoter is used to control the protein production. C–E) shows the production protocol for the three different scales used, small- medium- and pilot-scale respectively.

Fig. 2. Bioproduction of the human secretome in CHO cells in small-scale cultivations. A) The success rates for the CHO clones are shown. The overall success rate was almost 60%, although differing considerably between different groups based on predicted localization (Table 1), where proteins targeted for the digestive system are the group with the highest success. B) The amount of the different target proteins generated by the stream-lined approach; see Suppl. Table 2 for details regarding protein amounts. C) SDS-PAGE gels showing three representative proteins after purification: the first lane shows the molecular marker, the second lane shows purified protein sample and the third lane the same protein after deglycosylation. The enzymes from the deglycosylation kit are indicated by their respective names to the right.
Quality control of the produced and purified proteins

During the production process the quantity and biophysical quality of the target proteins produced were analyzed at different time points. The first analysis of the proteins from small- and medium-scale production took place 6 and 3 days after transfection, respectively. Supernatant for which the WB analysis of the culture media showed the expected protein band were regarded as having passed the analysis while proteins with a blank WB required further evaluation by analysis also of the cell lysate on WB. Where the WB was blank for conditioned medium as well as lysate, the protein was failed and classified as “no production”. Proteins that showed a band in the lysate lane but not in the lane for conditioned medium were clearly produced but not secreted and therefore failed and classified as “no secretion”. For the proteins that passed the first analysis, production continued until harvest at day 13. At harvest, an aliquot of the supernatant was analyzed on SDS-PAGE as well as WB. Proteins with a clear band of expected size proceeded to purification while proteins with a weak target band or showing < 80% of expected size were failed and classified as “low production” and “degradation”, respectively. Purity was confirmed using SDS-PAGE as well as WB including analysis of deglycosylated protein samples. The identity of the proteins that fulfilled the purity criterion of at least 80% of the sample having the expected size, was finally verified using MS/MS.

Reason for failure in the small-scale production pipeline

To understand if there was any common feature among the proteins that were difficult to produce, the possible connection was investigated between failure rate, failure reason and protein characteristics of size and hydrophobicity, as well as expected localization. All the proteins that were produced and purified passed through different quality assurance steps where they were sorted according to the analytic results (see above), i.e. passed or failed. Within the group of failed proteins there were five different classes according to the reason for failure, namely: degradation (218), low production (254), no production (134), no secretion (183) and a small group with inconsistent data, denoted ‘others’ (48). Regardless of annotated localization, all groups had proteins in all five fail-classes (Fig. 3). However, among the group of proteins that failed due to degradation, those annotated with a function that were produced and purified were analyzed at different time points. The first analysis of the proteins from small- and medium-scale production took place 6 and 3 days after transfection, respectively. Supernatant for which the WB analysis of the culture media showed the expected protein band were regarded as having passed the analysis while proteins with a blank WB required further evaluation by analysis also of the cell lysate on WB. Where the WB was blank for conditioned medium as well as lysate, the protein was failed and classified as “no production”. Proteins that showed a band in the lysate lane but not in the lane for conditioned medium were clearly produced but not secreted and therefore failed and classified as “no secretion”. For the proteins that passed the first analysis, production continued until harvest at day 13. At harvest, an aliquot of the supernatant was analyzed on SDS-PAGE as well as WB. Proteins with a clear band of expected size proceeded to purification while proteins with a weak target band or showing < 80% of expected size were failed and classified as “low production” and “degradation”, respectively. Purity was confirmed using SDS-PAGE as well as WB including analysis of deglycosylated protein samples. The identity of the proteins that fulfilled the purity criterion of at least 80% of the sample having the expected size, was finally verified using MS/MS.

Using human cell line HEK 293 to rescue difficult proteins

For proteins that failed production in CHO cells, the possibility that these could be rescued by expression in a human cell line was investigated. Some of the constructs previously failed due to degradation were therefore transfected into HEK 293 cells and transiently produced in a 4-day cultivation with a final culture volume of 40 mL. By using this protocol, up to 2.5 mg of pure protein was generated from a single culture. Many proteins that showed degradation in CHO cells were successfully produced in the human cell line, see examples in Fig. 4 (Supplementary Table S2). Out of 126 protein constructs that had earlier failed in the CHO cell factory due to degradation, 86 (68%) were successfully rescued by changing the production host from CHO to HEK 293. These results suggest that using a human cell line would be an attractive option for difficult to produce proteins in the standard bio-manufacturing CHO host cell line due to degradation. One possible explanation for why these proteins do not degrade in the human cell line is the shorter production protocol used. Furthermore, hosts of different origin with differences in protein processing machinery, may have different impacts on the degradation pattern.

Medium-scale production

To be able to meet the demand for larger protein amounts than were possible to generate in the small-scale production pipeline, a medium-scale pipeline was developed (Fig. 1D). All constructs selected for production at medium-scale had previously been successfully produced at a small-scale and to date 369 different proteins have been produced with an average yield of > 35 mg purified protein (Fig. 5A). For 86% of the proteins, > 10 mg purified product was achieved, with the highest amount being 152 mg achieved for proprotein convertase subtilisin/kexin type 9, PCSK9, an important target protein in immunotherapy for atherosclerosis [21]. When comparing the amount of pure protein from the different scales of production, it could be concluded that the medium-scale protocol generally performed better than the small-scale (Fig. 5B) without lowering the quality of the end product (Fig. 5C).

Pilot-scale production

To be able to increase further the amount of protein produced, a fed-batch protocol using WAVE Bioreactor systems was developed, aiming
for a final culture volume of 10 L (Fig. 1E). This protocol has to date been used for production of six different proteins, generating amounts of up to 1 g pure protein for antithrombin (SERPINC1). When comparing the amount of pure protein per volume cultivation from the three different scales, it is apparent that the protein production levels were highest with pilot-scale (Supplementary Table S3). Even though the number of proteins was rather small, their quality was similar to that achieved on a smaller scale (Fig. 6), and thus increasing the scale did not compromise the purity of the final product as assessed by SDS-page, WB and deglycosylation.

**Development of an MS-based quantification method**

To simultaneously determine concentration and identity of the produced proteins, an MS-based method for duplex serial absolute quantification was developed. This method is based on spike-in of stable isotope-labeled protein fragments (QPrESTs) corresponding to the target proteins [7]. The isotope-labeled QPrESTs are fused to a QTag, which is used for purification and subsequent quantification. Heavy labeled peptides coming from the QPrESTs are, together with a light quantification tag (QTag) of known concentration, used to enable serial quantification. First, the light peptides from the quantification tag are compared to the heavy labeled peptides from the QPrEST and thereafter the heavy labeled peptides from the QPrEST are compared with the corresponding light peptides from the produced protein. Thereby, both the amount of QPrEST and produced protein can be determined in a single MS experiment. A comparison between the targeted proteomics analysis and the determination of protein concentrations by spectrophotometry demonstrated that the latter more often generated over- or under-estimation of the concentration of the target protein (Fig. 7, Supplementary Table S4). Hence, it was decided to determine the absolute concentration of each protein using the MS based workflow and the production levels for each construct is presented in Supplementary Table S4.

**Use of the secretome resource to develop proteomics assays**

Data-independent acquisition (DIA) mass spectrometry can provide highly quantitative MS2-data for thousands of proteins in a single analysis, but requires libraries of peptide fragmentation spectra and retention times for identification of peptide peaks during the data analysis step [22,23]. These fragmentation spectra are often based on
shotgun proteomics analysis of complex samples which might result in low quality spectra where low abundant proteins are often missing. Therefore, it was decided to make use of the secretome resource to establish a high-quality spectral library by analyzing trypsin digested equimolar pools of the purified proteins in shotgun mode (Fig. 8A). In total, secreted proteins corresponding to 368 unique genes were screened and the fragmentation spectra were manually verified using the Skyline software [24]. For 340 of the proteins, DIA assays could be created based on at least one tryptic peptide, while 28 of the proteins did not result in any peptide fragmentation spectra. Human plasma samples were then analyzed and peptide data corresponding to 83 proteins could be extracted using the developed spectral library, while 257 proteins could not be detected using this DIA proteomics analysis (Fig. 8B). As expected, proteins that are actively secreted into the blood were detected in plasma to a larger extent than any other category.

Fig. 6. Bioproduction of the human secretome in CHO cells using WAVE bioreactors. SDS-PAGE gels showing the three proteins earlier produced in small and medium-scale (Figs. 2C and 5 C) after production and purification in pilot-scale, the first lane showing the molecular marker, the second lane showing the purified protein and the third lane the purified protein after a deglycosylation treatment. As can be seen, the quality of the proteins is similar regardless of the scale that had been used for the production step.

Fig. 7. Absolute quantification by MS of the produced proteins. By utilizing QPrESTs originating from the Human Protein Atlas[1] absolute quantifications of the produced proteins were made using LC-MS/MS. A). The histogram shows the difference between the absolute concentration determined by LC-MS/MS and the concentration obtained through absorbance measurements and B) the scatterplot, shows the ratio between spiked in QPrEST and protein product used for the quantification.

Fig. 8. Use of the secretome resource to set up a DIA analysis. A) Schematic view of the proteomics assay development. Proteins were combined into pools of eight and digested with trypsin before analysis in shotgun-MS mode. Peaks of detected peptides were inspected manually and their fragmentation spectra and retention times were used for data extraction through DIA. B) Beeswarm plot showing the secretome proteins and ECDs for which DIA assays could be established and whether they could be detected (red) or not (black) in human plasma analyzed through DIA. The proteins are divided into their respective secretome classification categories and the y-axis shows the maximum normalized mRNA expression (NX) in human normal tissue (according to http://www.proteinatlas.org).
although some belonging to the non-blood categories, especially those with high relative mRNA expression levels in at least one tissue, were detected in plasma most likely due to leakage from cells undergoing apoptosis. It is noteworthy that although many of the gastric proteins were found in plasma, some belonging to the non-blood categories, especially those of the human secretome, have recently been shown to re-enter cells to aid in the development of drug candidates as described previously [14,27–30]. The library of proteins generated in-house was used to explore the induction of dedifferentiation of the EndoC-βH1 cell line. Dedifferentiation was monitored by measuring changes in the subcellular location and expression of the β-cell marker MAFA (a key regulator of glucose stimulated-insulin secretion) and the pre-endocrine marker SOX9 (Fig. 9A). Interestingly, FGF9 showed a higher activity than the FGF2 positive control, but FGF1, FGF4 and FGF18 also affected differentiation state (Fig. 9B). As far as we are aware, the activity of several of these fibroblast growth factors has not been described before in this context. Single cell analysis of FGFR expression on primary human islets showed that FGFR1 was dominantly expressed (Fig. 9C). Taken together, the data suggest that therapeutic intervention could be achieved by inhibition of FGFR1 signaling with no additional signaling pathways identified that maintain the dedifferentiated state of β cells.

Conclusion

The human secretome is a highly interesting group of proteins both in studies of human biology and as targets for the development of new drugs and diagnostics. Here, we report a high-throughput mammalian cell factory for recombinant expression of the secretome in CHO and HEK 293 cells. 1,276 human proteins were produced, successfully purified and analyzed regarding concentration and purity. All the data from the mammalian cell factories are available to enable further explorations of factors important for successful bioproduction in CHO and/or human HEK 293 cell lines. This protein resource has been used both for generating a spectral library used to identify proteins in MS-based data independent acquisition (DIA) workflows aimed at analysis of human blood and for phenotypic assays involving β-cell dedifferentiation. The phenotypic assay generated several actives from the fibroblast growth factor family, and the proteomics assays were used to analyze the presence or absence of the target protein in human plasma.

Acknowledgments

Funding was provided from the Novo Nordisk Foundation, the Knut and Alice Wallenberg Foundation, AstraZeneca and Vinnova (2017-02105). We acknowledge the entire staff of the Human Protein Atlas program for their valuable contributions.

Appendix A. Supplementary data

Supplementary material related to this article can be found in the online version, at doi:https://doi.org/10.1016/j.nbt.2020.05.002.

References

[1] Uhlen M, Fagerberg L, Hallström BM, Lundskog C, Oksvold P, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347:1260419https://doi.org/10.1126/science.1260419.
[2] Uhlen M, Karlsson MJ, Hober A, Svensson AS, Scheffèl J, et al. The human secretome. Sci Signal 2019:12.https://doi.org/10.1126/scisignal.aaz0274.
[3] Altay G, Peterz B. Gene Regulatory Cross Networks: Inferring Gene Level Cell-to-Cell Communications of Immune Cells. bioRxiv 2018:415943https://doi.org/10.1101/415943.
[4] Dougherty WG, Cary SM, Parks TD. Molecular genetic analysis of a plant virus polyprotein cleavage site: a model. Virology 1989;171:356–64. https://doi.org/10.1016/0042-6822(89)90484-7.
[5] Stearns DJ, Kurosawa S, Sims PJ, Esmon NL, Esmon CT. The interaction of a Ca2+-dependent monoclonal antibody with the protein C activation peptide region. Evidence for obligatory Ca2+ binding to both antigen and antibody. J Biol Chem 1988;263:826–32.
Diedisheim M, Oshima M, Albagli O, Huldt CW, Ahlstedt I, et al. Modeling human pancreatic beta cell dedifferentiation in Human Type 2 Diabetes. J Clin Endocrinol Metab 2016;101:1044–54. https://doi.org/10.1210/jc.2015-2860.

Oshima M, Pechberty S, Bellini L, Gopel SO, Campana M, et al. Stearoyl CoA desaturase is a gatekeeper that protects human beta cells against lipotoxicity and maintains their identity. Diabetologia 2020;63:395–407. https://doi.org/10.1007/s00125-019-05046-x.

Lin H, Lee E, Hestir K, Leo C, Huang M, et al. Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. Science 2008;320:807–11. https://doi.org/10.1126/science.1154370.

Swiney DC, Anthony J. How were new medicines discovered? Nat Rev Drug Discov 2011;10:507–19. https://doi.org/10.1038/nrd3480.

Eder J, Sedran R, Wissmann C. The discovery of first-in-class drugs: origins and evolution. Nat Rev Drug Discov 2014;13:577–87. https://doi.org/10.1038/nrd4436.

Gonzalez R, Jennings LL, Knuth M, Orth AP, Kloek HE, et al. Screening the mammalian extracellular proteome for regulators of embryonic human stem cell pluripotency. Proc Natl Acad Sci U S A 2010;107:3552–7. https://doi.org/10.1073/pnas.0914019107.