The missing protein landscape of human chromosomes 2 and 14: progress and current status

Duek, Paula; Bairoch, Amos Marc; Gateau, Alain; Vandenbrouck, Yves; Lane, Lydie

How to cite

DUEK, Paula et al. The missing protein landscape of human chromosomes 2 and 14: progress and current status. In: Journal of proteome research, 2016, vol. 15, n° 11, p. 3971–3978. doi: 10.1021/acs.jproteome.6b00443

This publication URL: https://archive-ouverte.unige.ch/unige:86174

Publication DOI: 10.1021/acs.jproteome.6b00443

Perspective

The missing protein landscape of human chromosomes 2 and 14: progress and current status

Paula Duek, Amos Bairoch, Alain Gateau, Yves Vandenbrouck, and Lydie Lane

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.6b00443 • Publication Date (Web): 03 Aug 2016

The missing protein landscape of human chromosomes 2 and 14: progress and current status

Paula Duek[1], Amos Bairoch[1,2], Alain Gateau[1], Yves Vandenbrouck[3,4,5,] and Lydie Lane[1,2]*

1. *CALIPHO Group, SIB-Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland*
2. *Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, 1, rue Michel-Servet, 1211 Geneva 4, Switzerland*
3. *CEA, DRF, BIG, Laboratoire de Biologie à Grande Echelle, 17 rue des martyrs, Grenoble, F-38054, France*
4. *Inserm U1038, 17, rue des Martyrs, Grenoble F-38054, France*
5. *Université de Grenoble, Grenoble F-38054, France*

*: Corresponding author contact information: lydie.lane@sib.swiss tel: 00 41 22 379 58 41

Email addresses:

paula.duek@sib.swiss

yves.vandenbrouck@cea.fr

alain.gateau@sib.swiss

ab@sib.swiss

lydie.lane@sib.swiss

Keywords: human proteome project, missing proteins, mass spectrometry proteomics, bioinformatics, data mining, RNA sequencing

Abstract

Within the C-HPP, the French and Swiss teams are responsible for the annotation of proteins from chromosomes 14 and 2, respectively. neXtProt currently reports 1231 entries on chromosome 2 and 624 entries on chromosome 14; of these, 134 and 93 entries are still not experimentally validated and are thus considered as "missing proteins" (PE2-4), respectively. Among these entries, some may never be validated by conventional MS/MS approaches because of incompatible biochemical features. Others have already been validated, but are still awaiting annotation. Based on information retrieved from the literature and from three of the main C-HPP resources (Human Protein Atlas, Peptide Atlas and neXtProt), a subset of 40 theoretically detectable missing proteins (15 on chromosome 14 and 25 on chromosome 2) was defined for upcoming targeted studies in sperm samples. This list is proposed as a roadmap for the French and Swiss teams in the near future.

Introduction

The Chromosome-Centric Human Proteome Project (C-HPP) federates teams from different countries aiming at delivering an extended catalogue of experimentally validated human proteins[1]. Its first goal is to obtain definitive proof for the existence of at least one representative protein per protein coding gene by various approaches, including direct protein sequencing, antibody or mass-spectrometry based techniques. The ultimate goal is to elucidate the function of each protein, although increasing the throughput of functional studies remains challenging.

neXtProt[2] is a knowledgebase that collects information on human gene products from various resources at the genomic, transcriptomic and proteomic levels. The different products arising from one gene by alternative splicing or alternative initiation are generally grouped into a single entry. Based on the collected information, neXtProt assigns a "Protein Existence" (PE) score to each entry. A PE score of "1" means that at least one gene product described in the entry has been validated at the protein

level. A PE score of 2-4 is assigned to entries corresponding to gene products supported by genomic (PE3 and PE4) or transcriptomic (PE2) data, but awaiting experimental validation at the protein level. A PE5 score means that the corresponding gene has a low probability to encode a protein, based on available genomic or transcriptomic data. In the latest neXtProt release (2016-01-11), there are 2949 PE2-4 proteins out of a total of 20055 entries. One of the first aims of the C-HPP teams is to confidently detect these so-called "missing proteins" using mass spectrometry (MS) and antibody-based techniques[3]. Currently, a protein is considered validated by MS when two unique, non-nested peptides of at least 9 amino acids (aa) have been identified in human biological samples (www.thehpp.org/guidelines; Deutsch et al., 2016, submitted). MS data from the different C-HPP teams must be deposited via the ProteomeXchange system in order to be re-analyzed by PeptideAtlas through the Trans-Proteomic Pipeline[4] and integrated into neXtProt, which is now used as the reference knowledgebase for the project[5].

In the C-HPP context, the French and Swiss teams are responsible for the annotation of proteins from chromosomes 14 and 2, respectively. Over the past 3 years, they conducted a series of experiments combining shotgun MS/MS, single-reaction monitoring and immunohistochemistry to validate missing proteins in different organs or cell types[6,7]. Recently, they have been focusing on testis and sperm cells, which are expected to contain high numbers of missing proteins[8]. The datasets published in 2015 were submitted to ProteomeXchange and some of them were integrated into the 2016-01 PeptideAtlas build and subsequently into neXtProt release 2016-01-11, together with many other datasets from the C-HPP consortium.

neXtProt release 2016-01-11 reports 1231 entries on chromosome 2 and and 624 entries on chromosome 14, from which 18 and 17, respectively, may not correspond to genuine proteins and are flagged as PE5. However, there are still 134 entries on chromosome 2 and 93 entries on chromosome 14 that are still considered as "missing proteins" (PE2-4) (**Supplementary Table 1**). The aim of the present study was to select a subset of these missing proteins for future targeted MS studies, notably

on sperm samples, based both on sequence analysis and data mining in literature and C-HPP-linked resources. Our workflow, depicted in **Figure 1**, was composed of three steps : (i) discarding proteins that were experimentally validated but not annotated as such in neXtProt (ii) discarding the obvious difficult candidates (olfactory receptors, pseudogenes and proteins refractory to trypsin digestion) and (iii) prioritizing proteins with enriched expression in testis.

Defining a list of "theoretically detectable" missing proteins

Among the 227 missing proteins (PE2-4) on chromosomes 2 and 14, there are 45 proteins for which MS information is available in neXtProt (Supplementary Table 1, column J). These entries have not been upgraded to PE1 because this information does not comply with the current HPP requirements (Deutsch et al., 2016, submitted). They represent 25% of missing proteins on chromosome 2 (34 out of 134) and 12% of missing proteins on chromosome 14 (11 out of 93). Among them, nine have been unambiguously validated in our recent studies (Supplementary Table 1, in dark green): Two chromosome 2 proteins (TMEM169 and TEX261) have been unambiguously validated by targeted LC-SRM in glioblastoma cell lines[6] whereas three chromosome 14 proteins and four chromosome 2 proteins have been confirmed with several unique peptides of at least 9 aa by shotgun proteomics in sperm (Vandenbrouck et al, 2016, submitted) (Supplementary Table 1, column M). Notably, the three validated proteins on chromosome 14 (EDDM3A, ADAM21 and CATSPERB) have been suggested to play a role in the function or maturation of sperm[9–12]. For 16 protein entries (3 on chromosome 14 and 13 on chromosome 2), we provide publications that could be used by curators to confirm proteins' existence on the basis of orthogonal criteria such as functional assays or antibody-based techniques (Supplementary Table 1, in light green, column L). This list of publications is under examination by curators in light of the current criteria used to assign the PE1 score in UniProtKB/Swiss-Prot and neXtProt (www.uniprot.org/docs/pe_criteria). For the remaining 20 proteins (5 on chromosome 14 and 15 on chromosome 2), further experimental confirmation is needed (Supplementary Table 1, in yellow).

Among the 182 PE2-4 proteins for which no MS evidence is available in neXtProt (82 on chromosome 14 and 100 on chromosome 2), eighteen (10 on chromosome 14 and 8 on chromosome 2) have been confidently identified by several unique peptides in sperm (Vandenbrouck et al, 2016, submitted) or testis[13] (Supplementary Table 1, in dark green). Twenty-six (6 on chromosome 14 and 20 on chromosome 2) have related publications and might be upgraded to PE1 provided that the reported biochemical evidence meets the neXtProt/Swiss-Prot quality requirements (Supplementary Table 1, in light green). FAM71D (chromosome 14) and FER1L5 (chromosome 2) were detected by a single unique peptide of more than 9 aa in sperm and would need further confirmation by targeted assays (Supplementary Table 1, in yellow).

Taken together, the combination of MS information retrieved from neXtProt with our own datasets and data from the literature shows that 27 missing proteins were validated by more than 2 peptides (14 on chromosome 2 and 13 on chromosome 14) but not yet curated by PeptideAtlas and integrated into neXtProt, 22 were detected with a single peptide (16 on chromosome 2 and 6 on chromosome 14), and 42 have associated publications awaiting annotation (33 on chromosome 2 and 9 on chromosome 14) (Figure 1). This means that nearly half of the chromosome 2 missing proteins (63 out of 134) have been potentially already detected in human samples, whereas only 30% of the chromosome 14 missing proteins (28 out of 93) would have been detected. Hence, it seems that chromosome 14 missing proteins are less prone to detection than chromosome 2 missing proteins.

Among the 65 remaining missing proteins on chromosome 14, as many as 27 (42%) belong to the olfactory receptor family. These genes form a cluster located at 14q11.2. In contrast, there are only two olfactory receptors among the 71 remaining missing proteins on chromosome 2, located on 2q37.3. Olfactory receptors (Supplementary Table 1, in red) are notoriously difficult to detect because they are nearly exclusively expressed in a small subset of neurons located in a restricted region of the sensory epithelium. Interestingly, OR4N2 expression has been detected by RNA sequencing in testis (http://www.proteinatlas.org/ENSG00000176294-OR4N2/tissue), suggesting a potential expression in

non-chemosensory tissues, as previously described for a few other olfactory receptors (see[14] for review). To date, none of the 423 olfactory receptors encoded by the human genome has been reliably identified using MS-based techniques (E. Deutsch, personal communication). Identification of these proteins is one of the most challenging tasks for the C-HPP consortium as a whole and for the chromosome 14 team in particular. Two other proteins (GPR33 on 14q12 and GKN3P on 2p13.3) will probably be impossible to validate because they are encoded by inactivated genes (pseudogenes) in most human populations [15,16] (Supplementary Table 1, in grey).

Hence, there are 105 proteins on chromosomes 2 and 14 that have never been observed and are neither olfactory receptors nor pseudogenes. They represent 40% (37 out of 93) and 51% (68 out of 134) of the missing proteins on chromosome 14 and chromosome 2, respectively. To help design experimental protocols allowing the validation of these proteins by MS, we carefully examined their properties. Analysis of length distribution (column F) shows that these missing proteins are significantly shorter than proteins validated by 2 peptides of at least 9 aa (Kolmogorov-Smirnov test, p=0.001). Indeed, on chromosome 14, the mean length of this category of missing proteins is 399 aa (median 305 aa) whereas the mean length of the proteins validated by MS is 637 (median 456 aa). On chromosome 2, the mean length of this category of missing proteins is 468 aa (median 338 aa) whereas the mean length of the proteins validated by MS is 716 aa (median 493 aa) (data not shown). The smallest missing protein, C14orf144, is predicted as secreted with a signal peptide of 26 aa, which means that the mature protein would consist of only 28 aa. Fortunately, two theoretical tryptic proteotypic peptides of 10 and 14 aa can be found in SRMAtlas[17], meaning that this protein should be observable by MS provided it is expressed. Another small protein is COX8C, a mitochondrial protein whose predicted mature chain (after cleavage of potential transit peptide) would be 43 aa long. This sequence generates a single theoretical tryptic peptide of 32 aa harboring a transmembrane domain, a feature which is not optimal for MS detection.

Our first hypothesis was that, due to their smaller length, missing proteins might lack a sufficient number of detectable unique tryptic peptides. Thus, we computed the number of theoretical unique tryptic peptides of 9 - 50 aa for the canonical isoform of each missing protein entry (column K). To check the unicity of each peptide, we took into account the 2.5 million variants that are reported in neXtProt, but limited the combinations of variants to one variant per span of 6 aa, as described in Vandenbrouck et al, 2016, submitted. OTOS and C14orf132 have only one such theoretical unique tryptic peptide while LIMS3, WASH2P, POTEG and POTEM have none. Validation of these six proteins (Supplementary Table 1, in orange) would thus require specific and challenging protocols, notably digestion by enzymes other than trypsin. Before exploring the possibility of using other enzymes, we checked if we could find transcriptomic evidence for these proteins. We carefully examined the RNA sequencing data available on the Human Protein Atlas (HPA) website (version 15), coming from the analysis of 32 human tissues[8]. In this dataset, POTEG and POTEM could not be detected and LIMS3 was expressed only at low levels, suggesting that these three proteins will be difficult to detect in human samples, no matter which enzyme is used. No RNA sequencing information could be retrieved from the HPA website for WASH2P and C14orf132, whereas OTOS expression could be detected in thyroid and brain. We performed in silico digestion of OTOS, WASH2P and C14orf132 with various enzymes other than trypsin using the PeptideCutter tool on the ExPASy website (web.expasy.org/peptide_cutter/) and found that chymotrypsin would generate at least 2 unique peptides of 9 amino acids or more for C14orf132 and OTOS. In contrast, we were not able to find an enzyme that could generate unique peptides of less than 50 amino acids for WASH2P (data not shown).

The 99 other entries (Supplementary Table 1, in white) – 34 on chromosome 14 and 65 on chromosome 2 – have at least two theoretical unique tryptic peptides of [9-50 aa], making it theoretically possible to validate them by MS with respect to the current HPP guidelines. Notably, a

few of them have high numbers of transmembrane domains (column G) that may hinder their solubilization and thus their detection.

Prioritizing the theoretically detectable missing proteins that are present in sperm

One of the reasons these 99 proteins have escaped detection so far might reside in their spatially or temporally restricted expression pattern. To test this hypothesis, we examined the RNA sequencing data available on the Human Protein Atlas (HPA) website (version 15)[8]. Among the 99 "theoretically detectable" missing proteins, 76 have RNA sequencing information on the HPA website; 27 proteins (11 on chromosome 14 and 16 on chromosome 2) display a broad expression pattern (i.e. detected in 7 tissues or more), whereas 49 show a spatially restricted expression pattern (i.e. are detected in 6 tissues or less) (Supplementary Table 2). Most of these proteins seem to be expressed only at low level (< 10 FPKM), which will imply to develop specific enrichment or targeting procedures. Remarkably, 29 proteins (11 on chromosome 14 and 18 on chromosome 2) out of the 49 proteins with a restricted expression pattern were expressed only in testis or in a small group of tissues that include testis, indicating that they may be good candidates for targeted LC-SRM studies in sperm samples (Supplementary Table 2, column G, in green). The other 20 proteins have distinct tissue specificity (Supplementary Table 2, column G, in yellow). For example, GPR45 and C2orf80 are only expressed in brain, whereas SYNDIG1L and C2orf91 are only expressed in lung. This information can be used by the C-HPP consortium to look for these proteins in the appropriate biological samples.

Among the 27 proteins for which RNA sequencing information from HPA indicates a broad expression profile, eight are expressed at low levels (< 10 FPKM) in all the tissues studied, as well as in cell lines (data not shown). The detection of such proteins will probably be very difficult, implying specific enrichment procedures, for example by affinity purification. In contrast, ATP5G3 is highly expressed (>50 FPKM) in all tissues studied, as well as in most cell lines. It is an integral membrane protein which is predicted to be part of the mitochondrial membrane complex V. This protein can

probably be looked for in any cell, but its detection will need specific protocols for mitochondrial membrane preparation and protein solubilisation. Likewise, TMEM 37, TMEM253, TMEM178A, TMEM229B and SLC38A11 are expressed at least at medium levels (10-50 FPKM) in a number of tissues, but since they are integral membrane proteins, their detection will also probably require specific protocols for membrane preparation and protein solubilisation. KLF7 is a transcription factor which has been shown to be developmentally regulated in Xenopus and mouse [18,19], suggesting that it may also be developmentally regulated in human. Although it may be difficult to detect in adult tissues, we noticed that is expressed at medium levels in SH-SY5 neuroblastoma cells. Hence, it might be successfully detected in nuclear extracts from these cells. The twelve remaining proteins are widely expressed at low levels, but are expressed at higher levels in a restricted set of tissues. KLHL33 and KLHL30 seem to be enriched in skeletal muscle, KLHL28 in bone marrow, RPS6KL1 in cerebral cortex, ZNF514 in ovary and endometrium, and FAM178B in spleen. Interestingly, CCDC74B, MOK, C14orf79, EFCAB11, RGPD1 and RGPD3 are enriched in testis and/or fallopian tube (Supplementary Table 2, column G, in light green). They have been added to the list of proteins to be considered for studies in sperm samples (Supplementary Table 2, column I).

Presently neXtProt does not integrate RNA sequencing results, but integrates transcriptomics data derived from EST and microarray experiments after meta-analysis and quality scoring performed by the BGee[20] group, using a set of anatomical descriptors from CALOHA (available at ftp://ftp.nextprot.org/pub/current_release/controlled_vocabularies/caloha.obo). neXtProt also reports expression information that is extracted from published RT-PCR and Northern blot experiments by UniProtKB/Swiss-Prot curators. Fifty-four out of the 99 “theoretically detectable” missing proteins have high quality (Gold) transcriptomics data in neXtProt (Supplementary table 2, column H). Preferential expression in testis was clearly confirmed for 10 out of the 35 previously described candidates for targeted studies in sperm (Supplementary table 2, column I, in bold). These 10 proteins will be studied with high priority. For 11 others, a different profile was reported : four had a restricted

expression pattern outside testis, and seven had a broad expression profile. For the 14 other candidates, there was no available information in neXtProt.

We then examined available information about the 23 "missing proteins" for which no RNA sequencing information was available in HPA. Seven of them have high quality (Gold) expression data in neXtProt. According to microarray experiments, GBX2 is expressed in early stage embryo (Carnegie stage 2), thereby emphasizing its putative function as a transcription factor for cell pluripotency and differentiation. LINC01551 would be expressed in the nervous system, while DIRC1 would be expressed in vagina. Analysis of EST libraries indicates that PLGLA is expressed in liver, in agreement with the reported Northern blot data[21]. No high quality microarray or EST-based transcriptomics data could be retrieved for the orphan receptor GPR148 and the HERV-H_2q24.1 and HERV-H_2q24.3 provirus ancestral Env polyproteins, yet the three proteins were found to be expressed in testis by RT-PCR[22] and quantitative RT-PCR[23] and will be considered as additional candidates for targeted LC-SRM studies in sperm samples (Supplementary Table 2, column I).

Conclusion

This study highlights that among the 227 PE2-4 proteins encoded by genes on chromosomes 2 and 14, 99 are genuine missing proteins that have not been detected so far and represent suitable candidate for further investigation by applying our experimental workflows. We are confident that the information available in neXtProt and HPA will help us define the optimal conditions for their detection. We have been designing specific assays to validate the 38 proteins that are expected to be present in testis or sperm based on transcriptomics data. To this list, we will add two proteins (FAM71D and FER1L5) that were detected as single hits in sperm in our accompanying paper (Vandenbrouck et al., 2016, submitted). Interestingly, these two proteins were shown to be preferentially expressed in testis by RNA sequencing and/or microarray studies. In contrast, the 20 one hit wonders detected in other

studies do not seem to be preferentially expressed in testis. Our final selection of 40 proteins is shown in Table 1.

Acknowledgements

We thank Monique Zahn for critical reading of the manuscript. We deeply thank all our colleagues from the chromosome 2 and 14 teams for establishing solid experimental and bioinformatics workflows that meet C-HPP quality requirements. We thank the UniProt groups at SIB, EBI and PIR for their dedication in providing up-to-date high-quality annotations for the human proteins in UniProtKB/Swiss-Prot thus providing neXtProt with a solid foundation. We thank the PeptideAtlas, SRMAtlas, Human Protein Atlas and Bgee teams for openly sharing their data, tools and expertise with the community. neXtProt development benefits from extensive funding support from the SIB Swiss Institute of Bioinformatics. The neXtProt server is hosted by Vital-IT, the bioinformatics competence center that supports and collaborates with life scientists in Switzerland. This work was partially funded through the French National Agency for Research (ANR) (grant ANR-10-INBS-08; ProFI project, "Infrastructures Nationales en Biologie et Santé"; "Investissements d'Avenir" call).

References

(1) Paik, Y.-K.; Jeong, S.-K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H.-J.; Na, K.; Choi, E.-Y.; Yan, F.; et al. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221–223.

(2) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; et al. The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **2015**, *43* (D1), D764–D770.

(3) Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.* **2015**.

(4) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics. Clin. Appl.* **2015**, *9* (7-8), 745–754.

(5) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the human proteome project 2013-2014 and strategies for finding missing proteins. *J. Proteome Res.* **2014**, *13* (1), 15–20.

(6) Carapito, C.; Lane, L.; Benama, M.; Opsomer, A.; Mouton-Barbosa, E.; Garrigues, L.; Gonzalez de Peredo, A.; Burel, A.; Bruley, C.; Gateau, A.; et al. Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J. Proteome Res.* **2015**, *14* (9), 3621–3634.

(7) Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; et al. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, *14* (9), 3606–3620.

(8) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Tissue-based map of the human proteome. *Science (80-.).* **2015**, *347* (6220), 1260419–1260419.

(9) Yi, C.; Woo, J.M.; Han, C.; Oh, J.S.; Park, I.; Lee, B.; Jin, S.; Choi, H.; Kwon, J.T.; Cho, B.N.; Kim do, H.; Cho, C. Expression analysis of the Adam21 gene in mouse testis. *Gene Expr. Patterns 10* (2-3), 152–158.

(10) Dong, Y.; Pan, Y.; Wang, R.; Zhang, Z.; Xi, Q.; Liu, R.Z. Copy number variations in spermatogenic failure patients with chromosomal abnormalities and unexplained azoospermia. *Genet. Mol. Res.* **2015**, *14* (4), 16041–16049.

(11) Liu, J.; Xia, J.; Cho, K.-H.; Clapham, D. E.; Ren, D. CatSperbeta, a novel transmembrane protein in the CatSper channel complex. *J. Biol. Chem.* **2007**, *282* (26), 18945–18952.

(12) Kirchhoff, C.; Pera, I.; Rust, W.; Ivell, R. Major human epididymis-specific gene product, HE3, is the first representative of a novel gene family. *Mol. Reprod. Dev.* **1994**, *37* (2), 130–137.

(13) Zhang, Y.; Li, Q.; Wu, F.; Zhou, R.; Qi, Y.; Su, N.; Chen, L.; Xu, S.; Jiang, T.; Zhang, C.; et al. Tissue-Based Proteogenomics Reveals that Human Testis Endows Plentiful Missing Proteins. *J. Proteome Res.* **2015**, *14* (9), 3583–3594.

(14) Kang, N.; Koo, J. Olfactory receptors in non-chemosensory tissues. *BMB Rep.* **2012**, *45* (11), 612–622.

(15) Bohnekamp, J.; Böselt, I.; Saalbach, A.; Tönjes, A.; Kovacs, P.; Biebermann, H.; Manvelyan, H.M.; Polte, T.; Gasperikova, D.; Lkhagvasuren, S.; Baier, L.; Stumvoll, M.; Römpler, H.; Schöneberg, T. Involvement of the chemokine-like receptor GPR33 in innate immunity. *Biochem. Biophys. Res. Commun.* **2010**, *396* (2), 272–277.

(16) Geahlen, J. H.; Lapid, C.; Thorell, K.; Nikolskiy, I.; Huh, W. J.; Oates, E. L.; Lennerz, J. K. M.; Tian, X.; Weis, V. G.; Khurana, S. S.; et al. Evolution of the human gastrokine locus and confounding factors regarding the pseudogenicity of GKN3. *Physiol. Genomics* **2013**, *45* (15), 667–683.

(17) Kusebauch, U.; Deutsch, E. W.; Campbell, D. S.; Sun, Z.; Farrah, T.; Moritz, R. L. Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. *Curr. Protoc. Bioinformatics* **2014**, *46*, 13.25.1–28.

(18) Gao, Y.; Cao, Q.; Lu, L.; Zhang, X.; Zhang, Z.; Dong, X.; Jia, W.; Cao, Y. Kruppel-like factor family genes are expressed during Xenopus embryogenesis and involved in germ layer formation and body axis patterning. *Dev. Dyn.* **2015**, *244* (10), 1328–1346.

(19) Laub, F.; Aldabe, R.; Friedrich, V. Jr; Ohnishi, S.; Yoshida T. R. F. Developmental expression of mouse Krüppel-like transcription factor KLF7 suggests a potential role in neurogenesis. *Dev Biol.* **2001**, *233* ((2)), 305–318.

(20) Bastian, F.; Parmentier, G.; Roux, J.; Moretti, S. Bgee : Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *Data Integr. Life Sci.* **2008**, *5109*, 124–131.

(21) Lewis, V. O.; Gehrmann, M.; Weissbach, L.; Hyman, J. E.; Rielly, A.; Jones, D. G.; Llinás, M.; Schaller, J. Homologous plasminogen N-terminal and plasminogen-related gene A and B peptides. Characterization of cDNAs and recombinant fusion proteins. *Eur. J. Biochem.* **1999**, *259* (3), 618–625.

(22) Parmigiani, R. B.; Magalhães, G. S.; Galante, P. A. F.; Manzini, C. V. B.; Camargo, A. A.; Malnic, B. A novel human G protein-coupled receptor is over-expressed in prostate cancer. *Genet. Mol. Res.* **2004**, *3* (4), 521–531.

(23) De Parseval, N.; Lazar, V.; Casella, J.-F.; Benit, L.; Heidmann, T. Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. *J. Virol.* **2003**, *77* (19), 10414–10422.

(24) Takahashi, S.; Sakakibara, Y.; Mishiro, E.; Kouriki, H.; Nobe, R.; Kurogi, K.; Yasuda, S.; Liu, M.-C.; Suiko, M. Molecular cloning, expression and characterization of a novel mouse SULT6 cytosolic sulfotransferase. *J. Biochem.* **2009**, *146* (3), 399–405.

Legends

Figure 1: Selection of 40 candidate proteins for targeted experiments from the list of 134 PE2-4 entries on chromosomes 2 and 93 PE2-4 entries from chromosome 14. From the initial lists of missing proteins, we extracted lists of 65 and 34 " theoretically detectable missing proteins" on

chromosome 2 and 14, respectively, by discarding: olfactory receptors (red rectangle), putative proteins encoded by inactivated genes (pseudogenes) in most human populations (grey rectangle), proteins that are indistinguishable by trypsin-based MS/MS workflows (orange rectangle), proteins that were validated by our recent studies in respect to the C-HPP guidelines (dark green rectangle), proteins whose status needs to be revised based on published studies (light green rectangle) and "one hit wonder" proteins for which a single MS peptide is reported (yellow rectangle). By analysing transcriptomics datasets, we prioritized 38 of these proteins for targeted MS experiments in sperm samples (14 from chromosome 14 and 24 from chromosome 2). To this list, we added the 2 "one hit wonder" proteins that were observed in sperm (blue rectangle).

Table 1: Selected subset of 15 missing proteins on chromosome 14 and 25 missing proteins on chromosome 2 to be searched in sperm samples. The following information was retrieved from neXtProt release 2016-01-11: accession number (column A), chromosomal location (column B), gene and protein names (columns C and D). In column E are mentioned the single-hit identifications reported In Vandenbrouck et al. 2016 , submitted. Column F shows the RNA sequencing results retrieved from HPA (version 15). Affymetrix and EST data retrieved from neXtProt, as well as RT-PCR information found in the literature are reported in column G. In bold are proteins prioritized for assessment in the next future. N/A stands for no data available.

Supplementary Table 1: List of the 227 "missing proteins"(PE2-4) encoded on chromosomes 2 and 14, with associated information. The following information was retrieved from neXtProt release 2016-01-11: accession number (column A), protein existence (PE) level (column B), chromosomal location (column C), gene and protein names (columns D and E), length of the canonical isoform (column F), number of transmembrane (TM) or intramembrane (IM) segments on the canonical isoform (column G), annotated function (column H), associated EC number for enzymes (column I), availability of MS data in neXtProt (column J). The number of theoretical unique peptides on the canonical isoform is reported in column K. PubMed identifiers (PMID) for related publications are

reported in column L. Validation by MS/MS or SRM in recent studies is reported in column M. Row color code : In red: olfactory receptors. In grey: pseudogenes in most humans. In orange: proteins that cannot be validated by MS based on trypsin digestion. In dark green: proteins recently validated by the Swiss and French teams (Vandenbrouck et al, submitted, and [6]). In light green: proteins that may be upgraded to PE1 based on biochemical data retrieved from the literature. In yellow: proteins with unconfirmed MS evidence. In white: the set of the 99 theoretically detectable missing proteins.

Supplementary Table 2: List of the 99 theoretically detectable missing proteins on chromosomes 2 and 14, and selection of 38 proteins for studies in sperm samples based on their expression profile. The following information was retrieved from neXtProt release 2016-01-11: accession number (column A), protein existence (PE) level (column B), chromosomal location (column C), gene and protein names (columns D and E), length of the canonical isoform (column F). RNA sequencing results retrieved from HPA (version 15) are reported in column G. Affymetrix and EST data retrieved from neXtProt are reported in column H. Column I indicates the best candidates for targeted studies in testis or sperm samples. In bold are those confirmed both by RNA sequencing and microarray or EST data. Row color code : In dark green: proteins with restricted expression pattern including testis. In light green: proteins with broad expression pattern but enriched in testis. In yellow : proteins with restricted expression pattern but not expressed in testis. In light yellow : proteins with broad expression pattern which are enriched in another tissue than testis.

Table 1 : Selected subset of 15 missing proteins on chromosome 14 and 25 missing proteins on chromosome 2 to be searched in sperm samples

Accession	Chr. location	Gene name	Protein name	MS data	HPA RNAseq	other transcriptomics data
NX_A8MTL3	**14q11.2**	**RNF212B**	**RING finger protein 212B**	**N/A**	**Medium in kidney. Low in testis**	**Testis, oviduct, vagina, kidney, embryo (microarrays)**
NX_Q8TAA1	**14q11.2**	**RNASE11**	**Probable ribonuclease 11**	**N/A**	**Low in testis**	**Testis (EST)**
NX_Q8N9W8	**14q23.3**	**FAM71D**	**Protein FAM71D**	**1 peptide**	**Low in testis**	**Mouth, testis (microarrays)**
NX_O43506	**14q24.2**	**ADAM20**	**Disintegrin and metalloproteinase domain-containing protein 20**	**N/A**	**Low in testis**	**Mouth, testis, tendon (microarrays)**
NX_Q7Z4L0	**14q32.12**	**COX8C**	**Cytochrome c oxidase subunit 8C, mitochondrial**	**N/A**	**Medium in testis**	**Testis, oviduct, fetal ovary (microarrays)**
NX_Q8N9Y4	**14q32.12**	**FAM181A**	**Protein FAM181A**	**N/A**	**Medium in testis. Low in fallopian tube, cerebral cortex, thyroid gland, lung**	**Testis, hippocampus, oviduct, bronchus (microarrays)**
NX_A4IF30	14q22.3	SLC35F4	Solute carrier family 35 member F4	N/A	Low in prostate and testis	Broad
NX_Q9BUY7	14q32.11	EFCAB11	EF-hand calcium-binding domain-containing protein 11	N/A	Medium in testis, fallopian tube, thyroid gland. Low in urinary bladder, rectum, esophagus, kidney, tonsil, ovary, gallbladder, colon, endometrium, placenta, duodenum, prostate, smooth muscle, lymph node, cerebral cortex, adipose tissue, adrenal gland, stomach, appendix, small intestine, salivary gland, skin, bone marrow, lung, spleen, heart muscle, liver, pancreas	colon, skin, brain, medulla oblongata, hypothalamus, Subthalamic nucleus, corpus striatum, frontal lobe, hippocampus, occipital lobe, temporal lobe, midbrain, spinal cord, ovary, oviduct, endometrium, myometrium, vagina, bronchus, conjunctiva, retina, breast, fetal retina (microarrays)
NX_Q9P2D8	14q32.12	UNC79	Protein unc-79 homolog	N/A	Low in cerebral cortex, testis, adrenal gland, fallopian tube	Broad

NX_Q9UQ07	14q32.31	MOK	MAPK/MAK/MRK overlapping kinase	N/A	Medium in testis, fallopian tube. Low in thyroid gland, ovary, stomach, skin, endometrium, kidney, lung, gallbladder, urinary bladder, cerebral cortex, adrenal gland, smooth muscle, adipose tissue, heart muscle, prostate, spleen, esophagus, duodenum, salivary gland, small intestine, appendix, placenta, rectum, bone marrow, lymph node, colon, pancreas, tonsil, liver	gingiva, blood, heart atrium, skin, dermis, adrenal gland, ovary, pituitary gland, testis, thyroid, mammary gland, prostate, bone, cartilage, tendon, brain, Inferior olivari nucleus, Superior vestibular nuclei, Hypothalamus, Thalamus, Caudate nucleus, cerebral cortex, frontal lobe, hippocampus, parietal lobe, cerebellum, lateral ventricle, ovary, oviduct, endometrium, myometrium, vagina, vulva, epididimys, prostate, seminal vesicle, testis, lung, bronchus, nose, pleura, trachea, renal glomerus, urethra, eye, conjunctiva, retina, peritoneum, breast, mammary gland, adipose tissue, cartilage, tendon, embryonic cerebral cortex, embryonic liver, fetal telencephalon, fetal cerebral cortex, fetal retina, fetal kidney, fetal ovary, fetal testis (microarrays)
NX_Q96F83	14q32.33	C14orf79	Uncharacterized protein C14orf79	N/A	Medium in fallopian tube, testis. Low in thyroid gland, prostate, kidney, lung, skin, endometrium, gallbladder, ovary, stomach, cerebral cortex, adrenal gland, smooth muscle, urinary bladder, salivary gland, colon, rectum, esophagus, duodenum, spleen, pancreas, adipose tissue, placenta, appendix, lymph node,	mouth, parotid gland, skin, pituitary gland, testis, thyroid, prostate, cartilage, tendon, brain, Inferior olivari nucleus, amygdala, caudate nucleus, cerebral cortex, frontal lobe, hippocampus, parietal lobe, temporal lobe, oviduct, myometrium, endometrium,

					small intestine, heart muscle, tonsil, bone marrow	vagina, epididymis, seminal vesicle, testis, lung, bronchus, nose, pleura, eye, conjunctiva, parotid gland, breast, myometrium, fetal telencephalon, fetal cerebral cortex, fetal testis (microarrays)
NX_C9J3V5	14q32.33	TEX22	Testis-expressed sequence 22 protein	N/A	Low in testis	Cerebral cortex, fetal telencephalon (microarrays)
NX_Q6ZRR7	14q23.1	LRRC9	Leucine-rich repeat-containing protein 9	N/A	Low in testis and fallopian tube	N/A
NX_Q8N769	14q24.3	C14orf178	Uncharacterized protein C14orf178	N/A	Low in testis	N/A
NX_Q52M58	14q32.2	C14orf177	Putative uncharacterized protein C14orf177	N/A	Low in testis	N/A
NX_Q6IMI4	**2p23.3**	**SULT6B1**	**Sulfotransferase 6B1**	**N/A**	**Low in testis and fallopian tube**	**Kidney and testis** [24]
NX_Q8N7S2		**DNAJC5G**	**DnaJ homolog subfamily C member 5G**	**N/A**	**Medium in testis**	**Testis (microarrays)**
NX_Q8N6M5	**2p25.3**	**ALLC**	**Probable allantoicase**	**N/A**	**Low in testis**	**Liver, testis, brain (microarrays)**
NX_A0AVI2	**2q11.2**	**FER1L5**	**Fer-1-like protein 5**	**1 peptide**	**Low in testis.**	**Testis, skeletal muscle, oviduct, bronchus**
NX_Q56UN5	**2q21.3**	**MAP3K19**	**Mitogen-activated protein kinase kinase kinase 19**	**N/A**	**Medium in fallopian tube. Low in testis, lung, endometrium**	**Testis, corpus callosum, oviduct, lung, bronchus (microarrays)**
NX_A6NES4	**2q37.1**	**MROH2A**	**Maestro heat-like repeat-containing protein family member 2A**	**N/A**	**Low in testis, thyroid, kidney**	**Liver, parathyroid, testis (microarrays)**
NX_P0DJD0	2p11.2	RGPD1	RANBP2-like and GRIP domain-containing protein 1	N/A	Medium in testis. Low in placenta, bone marrow, stomach, tonsil, adrenal gland, lymph node, pancreas, appendix, duodenum, small intestine, thyroid gland, fallopian tube, liver, ovary, colon, lung, spleen, cerebral cortex, endometrium, esophagus, prostate, skin	N/A
NX_Q7Z4S9	2p11.2	SH2D6	SH2 domain-containing protein 6	N/A	Low in testis, duodenum, thyroid gland, small intestine, colon	Colon (microarrays)

NX_A6NKT7	2q12.2	RGPD3	RanBP2-like and GRIP domain-containing protein 3	N/A	Medium in testis. Low in thyroid gland, bone marrow, endometrium, placenta, ovary, fallopian tube, gallbladder, skin, urinary bladder, adrenal gland, lymph node, tonsil, cerebral cortex, prostate, smooth muscle, rectum, appendix, kidney, lung, adipose tissue, colon, esophagus, duodenum, small intestine, spleen, stomach, liver, heart muscle, salivary gland, pancreas	N/A
NX_Q0VF49	2q33.1	KIAA2012	Uncharacterized protein KIAA2012	N/A	Low expression in fallopian tube and testis	Oviduct, bronchus (microarrays)
NX_Q53R12	2q36.3	TM4SF20	Transmembrane 4 L6 family member 20	N/A	High in duodenum and small intestine. Low in colon, rectum, testis, smooth muscle	Broad
NX_Q6UX34	2q37.1	C2orf82	Uncharacterized protein C2orf82	N/A	Low in testis, liver, prostate	Broad
NX_A6NCI8	2p13.1	C2orf78	Uncharacterized protein C2orf78	N/A	Low in testis	N/A
NX_A8MVX0	2p22.1	ARHGEF33	Rho guanine nucleotide exchange factor 33	N/A	Low in testis, ovary, adrenal gland, cerebral cortex, endometrium, fallopian tube	not detected
NX_Q5MAI5	2p22.1	CDKL4	Cyclin-dependent kinase-like 4	N/A	Low in testis	N/A
NX_B5MCY1	2p24.1	TDRD15	Tudor domain-containing protein 15	N/A	Low in testis	N/A
NX_Q86YG4	2q14.1	NT5DC4	5'-nucleotidase domain-containing protein 4	N/A	Low in testis	not detected
NX_Q96LY2	2q21.1	CCDC74B	Coiled-coil domain-containing protein 74B	N/A	Medium in testis, fallopian tube. Low in endometrium, thyroid gland, prostate, ovary, adrenal gland, gallbladder, cerebral cortex, lung, smooth muscle, kidney, urinary bladder, skin, esophagus, lymph node	oviduct (microarrays)
NX_Q580R0	2q21.2	C2orf27A/B	Uncharacterized protein C2orf27	N/A	Low in testis and cerebral cortex	not detected
NX_A7E2S9	2q21.2	ANKRD30BL	Putative ankyrin repeat domain-containing protein 30B-like	N/A	Low in testis	N/A

NX_Q03828	2q31.1	EVX2	Homeobox even-skipped homolog protein 2	N/A	Low in testis, prostate, colon, rectum	N/A
NX_Q6UXQ4	2q33.1	C2orf66	Uncharacterized protein C2orf66	N/A	Low in adipose tissue, adrenal gland, testis	N/A
NX_Q8TDV2	2q21.1	GPR148	Probable G-protein coupled receptor 148	N/A	not detected	Testis (RT-PCR [22])
NX_Q9N2J8	2q24.1		HERV-H_2q24.1 provirus ancestral Env polyprotein	N/A	N/A	Testis (RT-PCR [23])
NX_Q9N2K0	2q24.3		HERV-H_2q24.3 provirus ancestral Env polyprotein	N/A	N/A	Testis (RT-PCR [23])

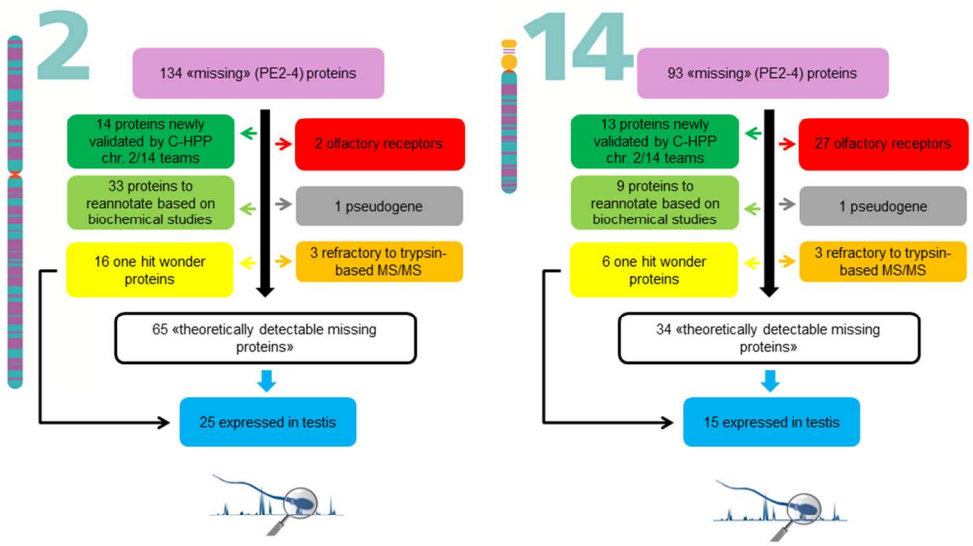
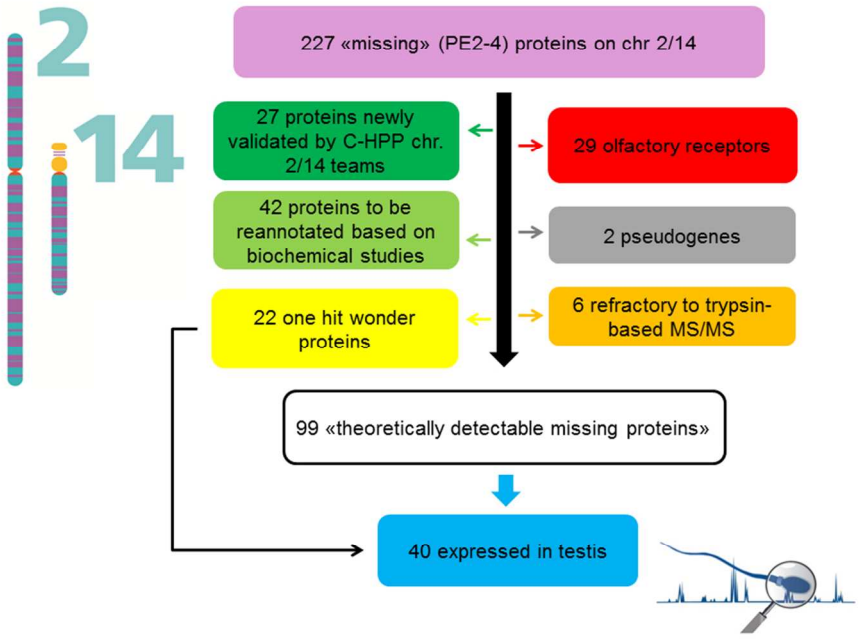


Figure 1: Selection of 40 candidate proteins for targeted experiments from the list of 134 PE2-4 entries on chromosomes 2 and 93 PE2-4 entries from chromosome 14. From the initial lists of missing proteins, we extracted lists of 65 and 34 " theoretically detectable missing proteins" on chromosome 2 and 14, respectively, by discarding: olfactory receptors (red rectangle), putative proteins encoded by inactivated genes (pseudogenes) in most human populations (grey rectangle), proteins that are indistinguishable by trypsin-based MS/MS workflows (orange rectangle), proteins that were validated by our recent studies in respect to the C-HPP guidelines (dark green rectangle), proteins whose status needs to be revised based on published studies (light green rectangle) and "one hit wonder" proteins for which a single MS peptide is reported (yellow rectangle). By analysing transcriptomics datasets, we prioritized 38 of these proteins for targeted MS experiments in sperm samples (14 from chromosome 14 and 24 from chromosome 2). To this list, we added the 2 "one hit wonder" proteins that were observed in sperm (blue rectangle).
Figure 1
275x190mm (96 x 96 DPI)



for TOC only

275x190mm (96 x 96 DPI)