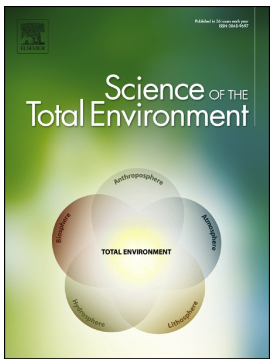
Journal Pre-proof



Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes

M. Carrascal, J. Abian, A. Ginebreda, D. Barceló

PII: S0048-9697(20)34674-X

DOI: https://doi.org/10.1016/j.scitotenv.2020.141145

Reference: STOTEN 141145

To appear in: *Science of the Total Environment*

Received date: 8 June 2020

Revised date: 16 July 2020

Accepted date: 19 July 2020

Please cite this article as: M. Carrascal, J. Abian, A. Ginebreda, et al., Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes, *Science of the Total Environment* (2018), https://doi.org/10.1016/j.scitotenv.2020.141145

Discovery of large molecules as new biomarkers in wastewater using environmental proteomics and suitable polymer probes

M. Carrascal[1], J. Abian[1*], A. Ginebreda[2], D. Barceló[2,3]

[1] Proteomics Laboratory CSIC/UAB, Institute of Biomedical Research of Barcelona, Spanish National Research Council (IIBB-CSIC/IDIBAPS), Rosellón 161, E-08036 Barcelona, Spain

[2] Institute of Environmental Assessment and Water Studies (IDAEA-CSIC), Department of Environmental Chemistry, Jordi Girona 18-26, 08034 Barcelona, Spain

[3] Catalan Institute for Water Research (ICRA), Emili Grahit 101, Parc Científic i Tecnològic de la Universitat de Girona, Edifici H2O, 17003 Girona, Spain

Corresponding author:

Joaquin Abian

Proteomics Laboratory CSIC/UAB. IIBB-CSIC. Rosellón 161 6a planta. E-08036 Barcelona, Spain.

e-mail: joaquim.abian.csic@uab.cat

phone: +35 93 5814853

ABSTRACT

The capability of monitoring large molecules as possible biomarkers in wastewater will be an important contribution to the new field of sewage epidemiology. Here, we explore the use of polymer probes together with untargeted proteomics for large scale protein analysis in sewage and treated water. Polymeric probes were immersed in the influent, anoxic reactor and effluent waters of a Spanish WWTP during 11 days. Proteins sorbed were extracted and identified by mass spectrometry. A total of 690 proteins from bacteria, plants and animals, including human, were identified showing different proteome profiles in the different sites. Bacterial proteins (510) pointed at 175 genera distributed in 22 bacteria classes. The most abundant were EF-Tu, GroEL and ATP synthase which were contributed by a high number of species. Human was the species contributing the greatest number of identified proteins (57), some in high abundance like keratins. Human proteins dominated in the influent water and were efficiently removed at the effluent. Several of the proteins identified (S100A8, uromodulin, defensins) are known disease biomarkers. This study provides the first insight into the proteome profiles present in real wastewater.

Keywords

Sewage epidemiology; sewage water; HPLC-HRMS; water fingerprinting, proteins; biomolecules.

1. INTRODUCTION

More than a decade ago, Christopher Wild coined the term “exposome” (Wild, 2005) to describe the entirety of human exposures, from both exogenous and endogenous sources, as an umbrella concept to help integrate the complexity of adverse health responses linked to exposures. Therefore, the exposome concept encompasses any environmental (i.e., non-genetic) factors including chemicals, diet components, physical factors, or psychosocial stressors, together with the corresponding biological effects resulting from them (Council, 2012; Lioy and Smith, 2013; Vermeulen et al., 2020). In that context, chemical factors have deserved particular attention. External xenobiotic (synthetic) chemicals potentially include 10^5-10^6 commonly used in household or industry (Vermeulen et al., 2020), together with their environmental or internal transformation products, while internal indicators can be reflected in endogenous biomarker changes (Daughton, 2018) . In that respect, the advances achieved by high-resolution mass spectrometry (HRMS), together with big-data treatment tools (chemometric, bioinformatics, databases) provide the suitable toolbox necessary to successfully cope with such a challenging analytical task (Schymanski et al., 2014).

On the other hand, the measurement of the exposure and effects related to the chemical exposome poses a major challenge regarding the representativity of the samples analyzed. Individual studies require a high number of cases to be statistically significant and become economically unaffordable. In that respect, sewage water has been proposed as an alternative approach, since it integrates the effluents of the entire population served by a certain wastewater system, hence providing an average picture of its health status (Daughton, 2018; Rice and Kasprzyk-Hordern, 2019). An additional advantage is that such kind of studies may be continued throughout time so that they can highlight possible changes occurred on the health status of the population monitored. Concepts like ‘sewage epidemiology’ (Daughton, 2011), ‘water fingerprinting’ (Rice and Kasprzyk-Hordern, 2019), ‘sewage chemical information

mining (SCIM)' (Daughton, 2018) or 'wastewater based epidemiology (WBE)' (Daughton, 2018) have been developed on this ground.

Whereas successful studies have been mostly carried out in the chemical exposure side, namely, the consumption of illegal drugs (Mastroianni et al., 2017; Thomas et al., 2012), pharmaceuticals and personal care products (Burgard et al., 2014; Gao et al., 2016), tobacco (Castiglioni et al., 2015; Ryu et al., 2016b) and alcohol use (Ryu et al., 2016a), pesticide exposure (Rousis et al., 2017), etc., and on genetic biomarkers (Ahmed et al., 2020; Yang et al., 2017), studies aimed at characterizing wastewater metabolomics and proteomics (i.e., the biological response side) are lacking. Even though several biomarkers have been proposed for that purpose, notably protein biomarkers, to the best of our knowledge, no systematic studies on wastewater proteomics have been published (Rice and Kasprzyk-Hordern, 2019). This is partly due to the inherent complexity and variability of sewage water, as well as, to the presence of other proteomes (bacteria, plants, animals, etc.) other than the human one. However, such non-human proteome sources present in wastewater should be regarded as an additional and valuable source of environmental information rather than a mere interference to the human proteome.

Up to now, most of the published papers on sewage epidemiology reported data only of small molecules. Alternatively, proteomic studies carried out at wastewater treatment plants (WWTPs) were focused exclusively on the microbial consortia proteome (Zhang et al., 2019). The present communication is, to our knowledge, the first study describing a simple and fast workflow suitable for the extraction of proteins from wastewater and their subsequent analysis using HRMS Orbitrap technologies. Owing to the complexity of wastewater, preliminary experiments for the direct protein extraction from the water presented diverse problems; however, it was conveniently achieved using polyester polymeric probes. Our preliminary results provide a first insight into the proteome profiles present in real wastewater. This paper provides a

new insight into the discovery of large molecules as new biomarkers in sewage, being a complementary diagnostic tool for epidemiologists.

2. MATERIAL AND METHODS

2.1 Polymeric devices preparation and water exposure

Polymeric devices were prepared using cap units (Exposmeter, Tavelsjo, Sweden) filled with polycaprolactonediol homopolymer (Polysciences, Warrington, USA) as described (Rivas et al., 2016). Polymeric devices were placed for 11 days in three different points at the Gavà-Viladecans (Barcelona, Spain) WWTP: influent water, anoxic reactor (denitrification) and effluent water (sites 1, 2 and 3, respectively. Figure S1).

2.2 Sample preparation and LC-MS/MS analysis.

Three devices per sampling site were processed. Slices from the polymer surface were cut and 500 µL of 2 % SDS, 50 mM DTT, 75 mM Tris-HCl were added. Samples were incubated for 30 min at 95 ºC with constant rotation and cooled for 45 min at 20 ºC. Supernatants were recovered and treated in a Bullet Blender (Casas et al., 2017). Then, samples were centrifuged and the supernatants were concentrated by SDS-PAGE (Paulo et al., 2013). Gels were stained with Coomassie and the band with the concentrated proteins was excised and digested with trypsin (Casanovas et al., 2009). Peptides were analyzed by HR-LC−MS/MS using an Agilent 1200 HPLC system coupled to an LTQ Orbitrap XL mass spectrometer (Thermo Scientific). Peptides were separated using a C18 pre-concentration cartridge (Agilent Technologies) connected to a C18 100 µm x 150 mm column (Nikkyo Technos Co, Tokyo, Japan). Separation was carried out at 400 nL/min using a 60-min linear gradient from 0 to 35 % solvent B

(Solvent A: H2O, 0.1 % (v/v) formic acid; solvent B: ACN, 0.1 % (v/v) formic acid). The LTQ Orbitrap XL was operated in data-dependent mode (Casanovas et al., 2017).

2.3 Data analysis

Raw data were processed using PEAKS (Zhang et al., 2012) against the UniProtKB/Swiss-Prot total database (release 2019_11). Search parameters were: trypsin, 0.02 Da precursor mass tolerance, 0.8 Da fragment mass tolerance. Matches were filtered for 0.1 % FDR at spectrum level. Information on the confidence of individual identifications is given in supplementary information (Table S3).

Jupyter notebooks (Perkel, 2018) (Python) were used to further analyze the PEAKS outputs in a traceable form. When protein groups were comprised of different strains of the same species, the first accession in the protein group was selected as protein head. In other cases, group heads were assigned based on the more represented species. Only proteins observed in at least 2 of the 9 samples (3 sites x 3 replicates) were considered for further analyses. Normalized Spectral Count (NSC), a variant of the Normalized Spectral Abundance Factor, was used as an indicator of the relative abundance of the proteins (McIlwain et al., 2012). NSC corresponds to the sum of the peptide sequence matches of all the peptides pointing to a protein normalized by the protein mass.

Detailed procedures for sections 2.1-2.3 are available in the Supporting Information (SI-Material and Methods).

3. RESULTS AND DISCUSSION

3.1 The wastewater proteome. Polymeric probes immersed in three different sites of the Gavà-Viladecans WWTP (entrance, exit and anoxic reactor) were recovered after 11 days of exposure. Influent water corresponds to sewages from several important

municipalities around the city of Barcelona. The average flow is 40.000 m^3/day and the amount of organic compounds in these waters correspond to a population equivalent of 375.000. Proteomic analysis of the material sorbed in the probes allowed the identification of 690 proteins. Using the sample treatment and purification methods developed, trypsin digestion and HR-LC-MS/MS, ion chromatograms showed no relevant interferences from the non-proteic organic matter present in sewage water. The use of polymeric probes overcomes the drawbacks related to the processing of high volumes of sewage water. Extraction of the proteins from the probes is effortless, uses low extraction volumes and could be easily automated if needed.

The collection of proteins identified consisted of a great variety of proteins from bacteria, plants, and animals including human (Table I, Table S1). The most represented taxonomic domain was that of bacteria (508 proteins), being *Escherichia coli*, *Desulfovibrio vulgaris* and *Acidovorax citrulli* the most frequent species. Among the bacterial proteins, elongation factor Tu (EF-Tu), 60 kDa chaperonin (GroEL) and ATP synthase were the most abundant (in terms of NSCs) and those that were contributed by a greater number of species. These proteins are among the most abundant in bacteria. EF-Tu, besides its well-known intracellular role during translation, has also diverse functions on the extracellular surface (Harvey et al., 2019). These functions include the interaction with membrane receptors and with extracellular matrix on the surface of plant and animal cells. The GroEL intervenes in the intracellular protein folding but also can act in intercellular signaling with different biological effects (Maguire et al., 2002). Finally, ATP synthase is a mitochondrial membrane protein responsible for the storage of energy in form of ATP.

The species with the highest contribution to the protein number was nevertheless *Homo sapiens* (9.9 %). Human proteins included hair- and skin-derived proteins (keratins, hornein, desmoplastin, and desmoglein), uromodulin, and pancreatic enzymes (chymotrypsin-like elastases, chymotrypsin C, phospholipase A2, and alpha-

amylase). Uromodulin is a kidney-derived protein which is the major protein excreted in urine. Uromodulin is an index of renal tubular function (Garimella and Sarnak, 2017) and has been proposed as a biomarker for sewage epidemiology (Rice and Kasprzyk-Hordern, 2019). Several human proteins with antimicrobial properties were also detected like defensines 3 and 5, which kill microbes by permeabilizing their plasma membrane, lysozyme C, with bacteriolytic functions, or the intelectins, which specifically recognize microbial carbohydrate chains. Other interesting proteins were those related to human immune response as immunoglobulins and the calcium- and zinc-binding protein S100A8, a component of calprotectin, which has a wide plethora of intra- and extracellular functions.

Table I: The most abundant human and bacterial proteins in the 3 sampling sites. For human, only proteins with unique peptides and NSC > 30 are considered. For bacteria, the 15 most abundant proteins (based on total NSC) are shown together with the number of bacterial genera contributing to each protein and the NCS for each site.

HUMAN							
Uniprot ID	Description	# peptides			NSC		
		Site 1	Site 2	Site 3	Site 1	Site 2	Site 3
-	Keratins	461	740	94	2903	5906	481
P09093	Chymotrypsin-like elastase family member 3A	22	15	-	788	288	-
Q99895	Chymotrypsin-C	16	7	-	240	78	-
P04054	Phospholipase A2	7	-	-	183	-	-
P05109	Protein S100-A8	2	2	-	92	36	-
O60844	Zymogen granule membrane protein 16	7	-	-	121	-	-
P59665	Neutrophil defensin 1	3	2	-	88	29	-
P59666	Neutrophil defensin 3	3	2	-	87	29	-
P61626	Lysozyme C	3	2	-	54	36	-
P0DOX7	Immunoglobulin kappa light chain	5	-	-	51	-	-
Q8WWU7	Intelectin-2	7	-	-	45	-	-
P04746	Pancreatic alpha-amylase	11	1		39	3	-
Q8WWA0	Intelectin-1	6	-	-	41	-	-
P01876	Immunoglobulin heavy constant alpha 1	5	3	-	23	15	-
P05090	Apolipoprotein D	4	1	-	35	2	-
P04745	Alpha-amylase 1	9	1	-	32	3	-
P08217	Chymotrypsin-like elastase family member 2A	5	-	-	32	-	-
BACTERIA							
Total NSC	Description	# bacterial genera			NSC		
		Site 1	Site 2	Site 3	Site 1	Site 2	Site 3
4758	Elongation factor Tu	38	52	21	1229	2356	1173
2746	60 kDa chaperonin	20	49	9	502	1998	246
1856	ATP synthase subunit alpha	14	42	18	283	974	599
1534	ATP synthase subunit beta	20	28	8	452	891	191
388	60 kDa chaperonin 2	6	10	1	91	288	9
388	60 kDa chaperonin 1	5	7	5	55	235	98
346	Elongation factor Tu 2	1	4	3	28	159	159
284	Hydroxylamine reductase	10	-	1	281	-	3
249	Elongation factor Tu 1	3	3	2	68	138	43
241	Adenylylsulfate reductase subunit alpha	1	1	1	162	78	1
223	Malate dehydrogenase	6	6	2	89	95	39
136	K(+)-insensitive PPi energized proton pump	4	4	-	71	65	-
128	30S ribosomal protein S14	-	-	1	-	-	128
127	30S ribosomal protein S10	-	-	1	-	-	127
125	Glyceraldehyde-3-phosphate dehydrogenase	1	1	4	1	72	52

Other species that were revealed to be present by specific proteins include mammals (mouse, cow, and rat), amphibians (*Xenopus laevis*), fish (*Danio rerio*), birds (*Gallus gallus*, *Anas platyrhynchos*) and plants (O*ryza sativa, Avena sativa, Prunus dulcis,* and *Solanum tuberosum)*. Overall, the more abundant proteins in the sewage waters analyzed were keratins (mainly from human skin but also from other mammals and bird feathers) and elastase.

It should be noted that protein and species assignations are subjected to the known limitations of bottom-up proteomic analysis and sequence databases. Reported species correspond to the best candidates pointed to by a set of database-matched spectra. For example, *Danio rerio* (zebrafish) is a model species for which we have a reference proteome with 3,152 proteins. The *Cyprinidae* family, to which *D. rerio* belongs, includes more than 1,200 species that due to their phylogenetic proximity probably show homology with those of zebrafish. Most of these species, however, have not been sequenced yet, and those that are, are only represented in the database with 265 proteins (data from Uniprot). Therefore, there is a very high probability that the identification of a protein from any of these species will be annotated as a *D. rerio* protein.

The presence of human and other non-bacterial proteins in the probes, some as major components, suggest that sorption of soluble proteins from water is the main mechanism for protein capture in the conditions employed. For bacterial proteins, we cannot discard additional contribution from cells and proteins in biofilms generated at the probe surface. Probably, most of these proteins arrive at the sampling probe with different modifications (hydrolysis, oxidation, deamidation, etc) either induced be the sewage environment or because they were excreted in that form. These modifications, especially protein chemical or enzymatic hydrolysis, could determine the half-life of the proteins in the sewage media and, in consequence, their detectability in the analytical process. However, protein hydrolysis would not be extensive as suggested by the

identification of only a small number of non-tryptic peptides (3 % when considering the 20 % more abundant peptides). This is also supported by the high sequence coverage of the more abundant proteins reported with peptides corresponding to different sections of the sequence as we show for the case of elastase (Figure S2). We did not detect proteins with masses below 8 kDa. The average mass of the reported proteins (55.5 kDa) was significantly higher than that of the proteins in the database (39.9 kDa). Differences were also observed between the average isoelectric points of the reported proteins and those in database (5.8 vs 7.1). The two pI distributions were clearly different showing an overrepresentation of acidic proteins in the identified collection. Solubility was also found slightly higher in our protein set than for the database (Figure S3). Further work is needed to characterize the type of proteins or proteins fragments surviving in the sewages, the mechanisms for peptide and protein sorption in the polymeric probes that determine their selectivity as well as the stability of these molecules in the probes along the exposition period.

3.2. Influent, effluent and anoxic reactor show different proteome profiles.

Different numbers of proteins were identified in the influent, the anoxic reactor and the effluent waters of the WWTP (337, 397 and 171 proteins, respectively. Table S1). Besides, the proteomes of these sites showed different protein profiles (Figure 2, Table S2). These differences allow the 3 sites to be distinguished based on their protein composition (see PCA analysis and hierarchic clustering, Figure S4 and S5). Proteins specific of each sampling site at the treatment plant are included in clusters 1, 2 and 3 (Figure 1).

The proteome of the influent is dominated by human proteins reflecting the urban location of the WWTP. Keratins and the chymotrypsin-like elastase family members 3A and 3B are the main proteins in this group. Elastase 3A had been detected in activated sludge and it is considered a significant constituent of urban wastewater (Kuhn et al.,

2011). Elastase is a stable, recalcitrant enzyme. The presence of undamaged elastase molecules in sewage waters is supported by the practically complete sequence coverage for this protein in our proteomic analysis and the tryptic character of its sequenced peptides (Figure S2). Thus, it cannot be discarded the presence of the functional enzyme in these waters. In fact, the characterization of elastase in the extracellular matrix of activated sludge flocs and digested sludge has raised concerns on possible environmental effects in the receiving waters (Westgate and Park, 2010). In our study, elastase, as well as the other confidently identified human proteins, except for keratins, shows diminished concentrations in the anoxic reactor and were undetectable in the effluent water. All these proteins classified in cluster 2 (Figure 1).

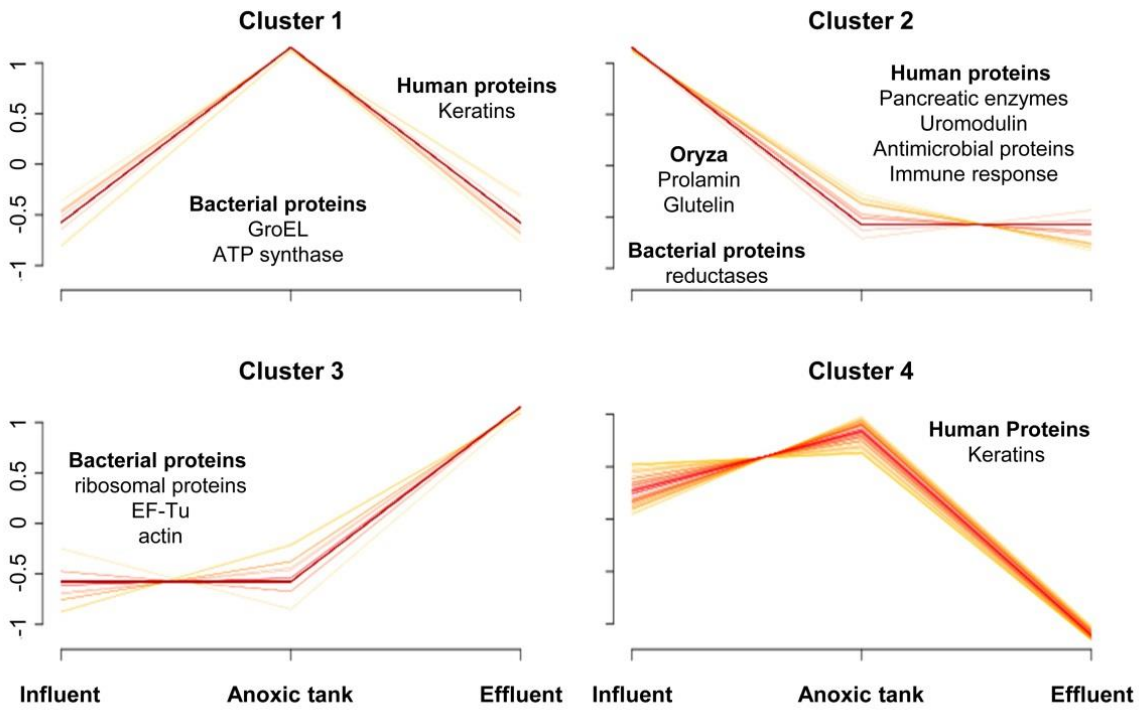


Figure 1.- Clustering of the proteins identified in the different collection sites (VSClust (Schwammle and Jensen, 2010)). Major members of each cluster are indicated. Clusters are ordered by the number of features (proteins) included. Clusters 1, 2, and 3 reveal site-specific proteins (anoxic tank, influent and effluent respectively.

Most keratins (90%) were classified at cluster 1 (22 keratins) or cluster 4 (55 keratins). In both cases, there is a concentration in the anoxic reactor of these proteins relative to the influent waters. This was probably due to their transport as part of the hair, feathers and skin small particles suspended in the sewage and being retained into the activated sludge of the anoxic reactor. Interestingly, other molecules strongly related to the skin structure such as those forming part of the desmosome (desmoglein, desmoplakin) or associated with keratin intermediate filaments (filaggrin) were also found concentrated in the anoxic reactor and following cluster 1 profile. All these proteins are effectively degraded or physically removed in the WWTP process as only a few of them could be detected in the effluent water and at much lower amounts than those measured at the anoxic reactor (10 % on average). Human keratins, although very diminished in comparison to the other sites, are still at relatively important components at the effluent. Nevertheless, bacterial proteins (actins, 30S ribosomal proteins, and EF-Tu), are the most abundant at this site, being most of them grouped in cluster 3. Although bacterial proteins were found in all the sampling sites, the dominant proteins in each site were different. Several reductases that were classified in cluster 2 were the most characteristic at the inlet. In contrast, GroEL and ATP synthase were the most characteristics of the anoxic reactor, probably reflecting the high bacterial activity on protein synthesis and energy production at this site. Finally, ribosomal proteins, EF-Tu and actin, mostly derived from *Acidovorax citrulli*, were the most abundant in the effluent water.

Profiles in figure 1 reflect the average composition at each site along the 11 days of exposition. The real-time dynamics of these traits is still to be studied. In this work, we used those long exposition times to accumulate as much material as possible to assure proteomics identification. Additional work is needed to optimize detection sensitivity allowing a higher resolution over time of composition changes.

3.3 Human proteins are eliminated in the WWTP.

A total of 57 human proteins were detected in the influent, some in high abundance as for the case of keratins. As it is shown for the 10 most abundant human proteins in the influent (excluding keratins) (Table I), all non-keratin proteins diminish in the anoxic reactor and are not detected in the effluent. For example, the 2 elastases taken together with Chymotrypsin account for a total of 1292 NSC in the influent but are reduced to one-third (434 NSC) in the anoxic reactor. Despite the high amounts of these proteins in the influent and the resistance of elastases to hydrolysis, they were not detected in the effluent.

3.4. Semiquantitative proteomics defines the different microbial proteomes in the sampling sites.

As shown above, the abundance of some protein species varies throughout the three sampling sites. In fact, the profiles of the bacterial proteins along the WWTP process could reflect both metabolic activity and diversity of bacterial communities acting at these sites (Figure 2, Figure S6). It should be noted that proteomic analyses, as performed here, could not only reveal the presence of proteins from more or less functional, viable cells but also of secreted proteins, or proteins from disrupted cells which are not necessarily physically present anymore in the sampling site.

The identified proteins pointed to 175 bacterial genera distributed in 22 bacterial classes (15, 18 and 10 in the influent, anoxic reactor and effluent, respectively). Species from the phylum Proteobacteria were the more abundant in the 3 sites, but with important relative differences between its classes. While Deltaproteobacteria was the most abundant class in the influent, its percentage diminished in the anoxic reactor and was close to zero in the effluent. The highest representative of this class at the influent site was *Desulfovibrio* that accounted for 41.8 % of all NSCs at this site. Contrarily, Betaproteobacterias which produce 84 % of all NSCs in the effluent (mostly

from *Acidovorax*) are only 10 % in the influent (mostly from *Azoarcus* and *Aromatoleum*). The most diverse bacterial population was found in the anoxic reactor where no single species stands out. The most abundant genera, *Burkholderia*, *Desulfovibrio*, and *Geobacter*, accounted together for only 14.7 % of the spectral counts.

Monitoring of enteric pathogens in water, such as *Escherichia* and *Salmonella*, is of interest for human health (Gorski et al., 2019). Together with other species of *Firmicutes* and *Bacteroidetes*, which are the dominant phyla in the human gut, they can reflect the level of water fecal contamination. Proteins from several species in these groups were present in the influent water and the anoxic reactor but none of them could be detected in the WWTP effluent.

3.5 Conclusions: Proteomics for wastewater-based epidemiology.

This communication highlights the capability of our approach for performing large scale proteomic analysis in sewage and treated water. While many works have focused on small molecules, metabolomic studies and microbial proteomics (Zhang et al., 2019), to the best of our knowledge, this is the first time the wastewater proteome is successfully and successfully assessed, and a significant number of human proteins are detected in this media.

Recently, the relevance of the profiling of human proteins in sewage waters for WBE has been stated by several authors (Choi et al., 2018; Daughton, 2012; Daughton, 2018; Mao et al., 2020; Rice and Kasprzyk-Hordern, 2019; Sims and Kasprzyk-Hordern, 2020). Rice et al. described this approach as a new paradigm for public health assessment. These authors proposed a group of urinary proteins including prostate specific antigen, C-reactive protein, interleukins 6 and 8, podocin, anterior gradient protein 2, and uromodulin (also reported here) as subjects of interest for community-wide biomonitoring of disease (Rice and Kasprzyk-Hordern, 2019). For their

part, Daughton et al., suggested vitamin D-binding protein, monocyte chemoattractant protein 1, nerve growth factor receptor, and gelsolin as candidate proteins for sewage biomonitoring (Daughton, 2018).

Identification in our work of human proteins such as uromodulin, α-amylase, and S100A8, which have been proposed as human health markers (Garimella and Sarnak, 2017; Mattes et al., 2014; Wang et al., 2018), opens new windows to this new field of sewage epidemiology. Such type of approach described here may be useful as well for proteomic profiling changes of different populations to better understand the scale of new epidemiological threats,.i.e Covid-19 (Daughton, 2020; Mao et al., 2020).

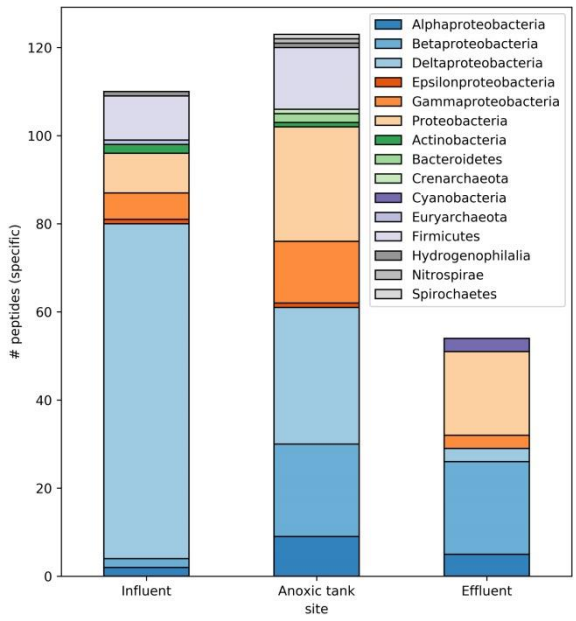


Figure 2.- Distribution of the bacterial phyla based on the number of specific peptides for each phylum (prepared with Unipept (Gurdeep Singh et al., 2019) web application). The phylum Proteobacteria is separated in its different Classes. The label Proteobacteria holds those peptide sequences that could not be assigned to a specific Proteobacteria class.

ACKNOWLEDGMENTS

We thank Vanessa Casas for technical assistance in sample preparation and mass spectrometric analysis, and Beatriz Reguera (Aigües de Barcelona) and the AMB (Àrea Metropolitana de Barcelona) for supporting in sampling collection. The Proteomics Laboratory CSIC/UAB is a member of Proteored-PRB3 and is supported by Grant PT17/0019/0008 of the PE I+D+I 2013-2016, funded by ISCIII and FEDER. This work was supported by the Spanish Ministry of Science and Innovation (Project CEX2018-000794-S).

REFERENCES

Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014; 8: 14; DOI: 10.3389/fninf.2014.00014.

Ahmed F, Tscharke B, O'Brien J, Thompson J, Samanipour S, Choi P, et al. Wastewater-based estimation of the prevalence of gout in Australia. Sci Total Environ 2020; 715: 136925; DOI: 10.1016/j.scitotenv.2020.136925.

Burgard DA, Banta-Green C, Field JA. Working upstream: how far can you go with sewage-based drug epidemiology? Environ Sci Technol 2014; 48: 1362-8; DOI: 10.1021/es4044648.

Casanovas A, Carrascal M, Abian J, Lopez-Tejero MD, Llobera M. Discovery of lipoprotein lipase pI isoforms and contributions to their characterization. J Proteomics 2009; 72: 1031-9; DOI: 10.1016/j.jprot.2009.06.002.

Casanovas A, Pinto-Llorente R, Carrascal M, Abian J. Large-Scale Filter-Aided Sample Preparation Method for the Analysis of the Ubiquitinome. Anal Chem 2017; 89: 3840-3846; DOI: 10.1021/acs.analchem.6b04804.

Casas V, Rodriguez-Asiain A, Pinto-Llorente R, Vadillo S, Carrascal M, Abian J. Brachyspira hyodysenteriae and B. pilosicoli Proteins Recognized by Sera of Challenged Pigs. Front Microbiol 2017; 8: 723; DOI: 10.3389/fmicb.2017.00723.

Castiglioni S, Senta I, Borsotti A, Davoli E, Zuccato E. A novel approach for monitoring tobacco use in local communities by wastewater analysis. Tob Control 2015; 24: 38-42; DOI: 10.1136/tobaccocontrol-2014-051553.

Council NR. Exposure Science in the 21st Century: A Vision and a Strategy. Washington, DC: The National Academies Press, 2012.

Choi PM, O'Brien JW, Li J, Jiang G, Thomas KV, Mueller JF. Population histamine burden assessed using wastewater-based epidemiology: The association of 1,4-methylimidazole acetic acid and fexofenadine. Environ Int 2018; 120: 172-180; DOI: 10.1016/j.envint.2018.08.009.

Daughton C. The international imperative to rapidly and inexpensively monitor community-wide Covid-19 infection status and trends. Sci Total Environ 2020; DOI: 10.1016/j.scitotenv.2020.138149.

Daughton CG. Illicit drugs: contaminants in the environment and utility in forensic epidemiology. Rev Environ Contam Toxicol 2011; 210: 59-110; DOI: 10.1007/978-1-4419-7615-4_3.

Daughton CG. Using biomarkers in sewage to monitor community-wide human health: isoprostanes as conceptual prototype. Sci Total Environ 2012; 424: 16-38; DOI: 10.1016/j.scitotenv.2012.02.038.

Daughton CG. Monitoring wastewater for assessing community health: Sewage Chemical-Information Mining (SCIM). Sci Total Environ 2018; 619-620: 748-764; DOI: 10.1016/j.scitotenv.2017.11.102.

Gao J, O'Brien J, Du P, Li X, Ort C, Mueller JF, et al. Measuring selected PPCPs in wastewater to estimate the population in different cities in China. Sci Total Environ 2016; 568: 164-170; DOI: 10.1016/j.scitotenv.2016.05.216.

Garimella PS, Sarnak MJ. Uromodulin in kidney health and disease. Curr Opin Nephrol Hypertens 2017; 26: 136-142; DOI: 10.1097/MNH.0000000000000299.

Gorski L, Rivadeneira P, Cooley MB. New strategies for the enumeration of enteric pathogens in water. Environ Microbiol Rep 2019; 11: 765-776; DOI: 10.1111/1758-2229.12786.

Gurdeep Singh R, Tanca A, Palomba A, Van der Jeugt F, Verschaffelt P, Uzzau S, et al. Unipept 4.0: Functional Analysis of Metaproteome Data. J Proteome Res 2019; 18: 606-615; DOI: 10.1021/acs.jproteome.8b00716.

Harvey KL, Jarocki VM, Charles IG, Djordjevic SP. The Diverse Functional Roles of Elongation Factor Tu (EF-Tu) in Microbial Pathogenesis. Front Microbiol 2019; 10: 2351; DOI: 10.3389/fmicb.2019.02351.

Kuhn R, Benndorf D, Rapp E, Reichl U, Palese LL, Pollice A. Metaproteome analysis of sewage sludge from membrane bioreactors. Proteomics 2011; 11: 2738-44; DOI: 10.1002/pmic.201000590.

Lioy PJ, Smith KR. A discussion of exposure science in the 21st century: a vision and a strategy. Environ Health Perspect 2013; 121: 405-9; DOI: 10.1289/ehp.1206170.

Maguire M, Coates AR, Henderson B. Chaperonin 60 unfolds its secrets of cellular communication. Cell Stress Chaperon 2002; 7: 317-29; DOI: 10.1379/1466-1268(2002)007<0317:cuisoc>2.0.co;2.

Mao K, Zhang K, Du W, Ali W, Feng X, Zhang H. The potential of wastewater-based epidemiology as surveillance and early warning of infectious disease outbreaks. Curr Opin Environ Sci Health 2020; DOI: 10.1016/j.coesh.2020.04.006.

Mastroianni N, Lopez-Garcia E, Postigo C, Barcelo D, Lopez de Alda M. Five-year monitoring of 19 illicit and legal substances of abuse at the inlet of a wastewater treatment plant in Barcelona (NE Spain) and estimation of drug consumption patterns and trends. Sci Total Environ 2017; 609: 916-926; DOI: 10.1016/j.scitotenv.2017.07.126.

Mattes W, Yang X, Orr MS, Richter P, Mendrick DL. Biomarkers of tobacco smoke exposure. Adv Clin Chem 2014; 67: 1-45; DOI: 10.1016/bs.acc.2014.09.001.

McIlwain S, Mathews M, Bereman MS, Rubel EW, MacCoss MJ, Noble WS. Estimating relative abundances of proteins from shotgun proteomics data. BMC Bioinformatics 2012; 13: 308; DOI: 10.1186/1471-2105-13-308.

Paulo JA, Kadiyala V, Brizard S, Banks PA, Conwell DL, Steen H. Short Gel, Long Gradient Liquid Chromatography Tandem Mass Spectrometry to Discover Urinary Biomarkers of Chronic Pancreatitis. Open Proteomics J 2013; 6: 1-13; DOI: 10.2174/1875039701306010001.

Perkel JM. Why Jupyter is data scientists' computational notebook of choice. Nature 2018; 563: 145-146; DOI: 10.1038/d41586-018-07196-1.

Rice J, Kasprzyk-Hordern B. A new paradigm in public health assessment: Water fingerprinting for protein markers of public health using mass spectrometry,. TRAC-Trend Anal Chem 2019; 119; DOI: 10.1016/j.trac.2019.115621.

Rivas D, Ginebreda A, Perez S, Quero C, Barcelo D. MALDI-TOF MS Imaging evidences spatial differences in the degradation of solid polycaprolactone diol in water under aerobic and denitrifying conditions. Sci Total Environ 2016; 566-567: 27-33; DOI: 10.1016/j.scitotenv.2016.05.090.

Rousis NI, Gracia-Lor E, Zuccato E, Bade R, Baz-Lomba JA, Castrignano E, et al. Wastewater-based epidemiology to assess pan-European pesticide exposure. Water Res 2017; 121: 270-279; DOI: 10.1016/j.watres.2017.05.044.

Ryu Y, Barcelo D, Barron LP, Bijlsma L, Castiglioni S, de Voogt P, et al. Comparative measurement and quantitative risk assessment of alcohol consumption through wastewater-based epidemiology: An international study in 20 cities. Sci Total Environ 2016a; 565: 977-983; DOI: 10.1016/j.scitotenv.2016.04.138.

Ryu Y, Gracia-Lor E, Bade R, Baz-Lomba JA, Bramness JG, Castiglioni S, et al. Increased levels of the oxidative stress biomarker 8-iso-prostaglandin F2alpha in

wastewater associated with tobacco use. Sci Rep 2016b; 6: 39055; DOI: 10.1038/srep39055.

Schwammle V, Jensen ON. A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. Bioinformatics 2010; 26: 2841-8; DOI: 10.1093/bioinformatics/btq534.

Schymanski EL, Singer HP, Longree P, Loos M, Ruff M, Stravs MA, et al. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. Environ Sci Technol 2014; 48: 1811-8; DOI: 10.1021/es4044374.

Sims N, Kasprzyk-Hordern B. Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level. Environ Int 2020; 139: 105689; DOI: 10.1016/j.envint.2020.105689.

Thomas KV, Bijlsma L, Castiglioni S, Covaci A, Emke E, Grabic R, et al. Comparing illicit drug use in 19 European cities through sewage analysis. Sci Total Environ 2012; 432: 432-9; DOI: 10.1016/j.scitotenv.2012.06.069.

Vermeulen R, Schymanski EL, Barabasi AL, Miller GW. The exposome and health: Where chemistry meets biology. Science 2020; 367: 392-396; DOI: 10.1126/science.aay3164.

Wang S, Song R, Wang Z, Jing Z, Wang S, Ma J. S100A8/A9 in Inflammation. Front Immunol 2018; 9: 1298; DOI: 10.3389/fimmu.2018.01298.

Westgate PJ, Park C. Evaluation of proteins and organic nitrogen in wastewater treatment effluents. Environ Sci Technol 2010; 44: 5352-7; DOI: 10.1021/es100244s.

Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005; 14: 1847-50; DOI: 10.1158/1055-9965.EPI-05-0456.

Yang Z, Xu G, Reboud J, Kasprzyk-Hordern B, Cooper JM. Monitoring Genetic Population Biomarkers for Wastewater-Based Epidemiology. Anal Chem 2017; 89: 9941-9945; DOI: 10.1021/acs.analchem.7b02257.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 2012; 11: M111 010587; DOI: 10.1074/mcp.M111.010587.

Zhang P, Zhu J, Xu XY, Qing TP, Dai YZ, Feng B. Identification and function of extracellular protein in wastewater treatment using proteomic approaches: A minireview. J Environ Manage 2019; 233: 24-29; DOI: 10.1016/j.jenvman.2018.12.028.

Table I: The most abundant human and bacterial proteins in the 3 sampling sites. For human, only proteins with unique peptides and NSC > 30 are considered. For bacteria, the 15 most abundant proteins (based on total NSC) are shown together with the number of bacterial genera contributing to each protein and the NCS for each site.

HUMAN

Uniprot ID	Description	# peptides			# NSC		
		Site 1	Site 2	Site 3	Site 1	Site 2	Site 3
	Keratins	461	740	94	2903	5906	481
P09093	Chymotrypsin-like elastase family member 3A	22	15	-	788	288	-
P08861	Chymotrypsin-like elastase family member 3B	11	5	-	264	68	-
Q99895	Chymotrypsin-C	16	7	-	240	78	-
P04054	Phospholipase A2	7	-	-	183	-	-
P05109	Protein S100-A8	2	2	-	92	36	-
O60844	Zymogen granule membrane protein 16	7	-	-	121	-	-
P59665	Neutrophil defensin 1	3	2	-	88	29	-
P59666	Neutrophil defensin 3	3	2	-	87	29	-
P63261	Actin cytoplasmic 2	8	5	13	19	11	77
P61626	Lysozyme C	3	2	-	54	36	-

BACTERIA

Uniprot ID	Description	# peptides			# NSC		
		Site 1	Site 2	Site 3	Site 1	Site 2	Site 3
T2G6Z9	Adenylylsulfate reductase|*Desulfovibrio gigas*	35	21	2	162	78	1
A1TJ05	Elongation factor Tu|*Acidovorax citrulli*	-	11	14	-	84	111
Q3JMP6	Elongation factor Tu|*Burkholderia pseudomallei*	-	11	13	-	88	103
Q62GK3	Elongation factor Tu|*Burkholderia mallei*	-	11	13	-	88	103
A1KB29	Elongation factor Tu|*Azoarcus sp.*	9	12	12	47	81	60
P42481	Elongation factor Tu|*Thiomonas delicata*	-	11	13	-	82	103
A2SCV1	60 kDa chaperonin|*Methylibium petroleiphilum*	-	18	11	-	83	60
Q2YAZ9	Elongation factor Tu|*Nitrosospira multiformis*	4	7	9	17	65	60
P10982	Actin-1 (Fragment)|*Absidia glauca*	-	-	10	-	-	133
A1W321	30S ribosomal protein S14|*Acidovorax sp.*	-	-	8	-	-	128

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Sample CRediT author statement

Montserrat Carrascal: Conceptualization, Methodology , Investigation, Writing - Original Draft. Joaquin Abian: Conceptualization, Software , Data Curation, Writing - Original Draft. Antonio Ginebreda: Conceptualization, Writing - review & editing. Damia Barceló: Conceptualization, Writing - review & editing, Project administration.

HIGHLIGHTS

- Large-scale proteomic analysis of sewage and treated water.
- Proteomics for Wastewater Based Epidemiology.
- Profiling of human proteins in sewage waters for WBE.
- Human uromodulin, α-amylase, and S100A8 identified in urban sewage.