Volume 14 | Issue 1 | April 2024 ISSN 2182-0287

Journal of Integrated OMICS

a methodological journal

JIOMICS

Journal of Integrated OMICS

Focus and Scope

Journal of Integrated OMICS, JIOMICS, provides a forum for the publication of original research papers, preliminary communications, technical notes and critical reviews in all branches of pure and applied "-omics", such as genomics, proteomics, lipidomics, metabolomics or metallomics. The manuscripts must address methodological development. Contributions are evaluated based on established guidelines, including the fundamental nature of the study, scientific novelty, and substantial improvement or advantage over existing technology or method. Original research papers on fundamental studies, and novel sensor and instrumentation development, are especially encouraged. It is expected that improvements will also be demonstrated within the context of (or with regard to) a specific biological question; ability to promote the analysis of molecular mechanisms is of particular interest. Novel or improved applications in areas such as clinical, medicinal and biological chemistry, environmental analysis, pharmacology and materials science and engineering are welcome.

JOURNAL OF INTEGRATED OMICS

A methodological Journal

ORIGINAL ARTICLES

ORIGINAL ARTICLES

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ARTICLE | DOI: 10.5584/jiomics.v14i1.228

Navigating the molecular diversity of SARS-CoV-2: early pandemic insights from comparative phylogenetic analysis.

Chahinez Triqui[1], Ilyes Zatla[2], Wafaa Lemerini[3], Nora Benmadani[4], Mohammed Charaf Eddine Houari[4], Semir Bechir Suheil Gaouar[4], Nassima Mokhtari-Soulimane[1]

[1]Laboratory of Physiology, Physiopathology and Biochemistry of Nutrition, Department of Biology, Faculty of Natural and Life Sciences, Earth and Universe, University of Tlemcen, Algeria. [2]Laboratory of Microbiology applied to the Food industry, Biomedical and the Environment, Faculty of Natural and Life Sciences, Earth and Universe Sciences. Department of Biology. University of Tlemcen, Algeria. [3]Laboratory of Organic Chemistry, Natural Substances, and Analysis, Faculty of Natural and Life Sciences, Earth and Universe Sciences. Department of Biology. University of Tlemcen, Algeria. [4]Laboratory of Applied genetic in agriculture, ecology and public health, Faculty of Natural and Life Sciences, Earth and Universe, University of Tlemcen, Algeria.

Available Online: 15 April 2024

Abstract

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in December 2019 precipitated the onset of the COVID-19 pandemic, which swiftly spread across more than 214 countries and territories, posing a significant global health crisis. In response, laboratories worldwide have embarked on extensive efforts to characterize the genomic landscape of the virus, employing a myriad of sophisticated genomic analysis techniques. This study endeavors to undertake a comprehensive exploration into the genetic diversity, geographical distribution, and virulence determinants of SARS-CoV-2 clades across 11 diverse countries, employing advanced computational biology methodologies. Leveraging molecular data sourced from prominent international databases, the analysis aims to unravel the intricate phylogenetic relationships and mutational dynamics exhibited by various viral strains circulating worldwide. The findings of this investigation promise to yield invaluable insights into the evolutionary trajectory of SARS-CoV-2, shedding light on potential therapeutic targets and informing strategies for mitigating the impact of the ongoing pandemic on global public health. Results highlight significant genetic diversity among SARS-CoV-2 strains across different countries, with phylogenetic analysis revealing distinct subclass groupings within each country. A manual comparison of sequences identified numerous mutations, with certain mutations associated with increased virulence. Comparison of clade G and clade O sequences revealed differences in mutation profiles, suggesting potential links to virulence and transmissibility. These findings underscore the dynamic nature of SARS-CoV-2 evolution and the importance of monitoring genetic changes for public health interventions.

Keywords: SARS-CoV-2, COVID-19, Mutation, Variant, Phylogeny.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel member of the Coronaviridae family, emerged as a global health threat in December 2019, originating from the bustling city of Wuhan, China. Its rapid transmission and virulence swiftly propelled the onset of the COVID-19 pandemic, spreading relentlessly across continents and precipitating unprecedented challenges to public health systems worldwide [1-4]. Notably, SARS -CoV-2 shares genetic similarities with other members of the coronavirus family, including Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [5-8]. Despite this genetic kinship, SARS-CoV-2 exhibits a distinctive clinical profile, encompassing a broad spectrum of manifestations ranging from mild respiratory symptoms to severe pneumonia and acute respiratory distress syndrome (ARDS). The multifaceted clinical presentation of COVID-19 underscores the imperative for a nuanced understanding of the genetic determinants underlying its pathogenesis and transmission dynamics [9-12].

Central to unraveling the complexities of SARS-CoV-2 infection is an elucidation of its genetic diversity and geographical distribution. The genetic variability exhibited by different SARS-CoV-2 strains holds pivotal implications for tracking transmission patterns, elucidating disease dynamics, and identifying potential therapeutic

***Corresponding author:** Ilyes Zatla, ilyes.zatla@univ-tlemcen.dz, Tel: +213-5403-15422

targets [13]. Against this backdrop, this study endeavors to undertake a comprehensive analysis of the genomic sequences of SARS-CoV-2 strains sourced from diverse geographical locations. By harnessing advanced computational biology methodologies, the investigation seeks to delineate the evolutionary dynamics and virulence determinants inherent to distinct viral lineages.

Specifically, our research provides valuable insights into the genetic diversity and mutational dynamics of SARS-CoV-2 strains, which are critical factors influencing the virus's antigenic properties and its ability to evade host immune responses. By characterizing the genomic evolution of the virus and identifying potential virulence determinants, our study contributes to the broader understanding of how SARS-CoV-2 interacts with the immune system and how these interactions may impact disease severity, transmission dynamics, and vaccine efficacy. Moreover, the identification of distinct viral clades and mutation profiles across different geographic regions underscores the importance of ongoing surveillance and molecular epidemiology efforts, which are essential for guiding immunization strategies and vaccine design initiatives. By elucidating the evolutionary trajectory of SARS-CoV-2, our research provides valuable insights that are pertinent to the field of immunology and have implications for public health interventions aimed at controlling the COVID-19 pandemic.

Materials and methods

Genome sequences of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) originating from 11 countries (Algeria, Germany, Australia, England, Spain, France, Italy, Saudi Arabia, Kuwait, Switzerland, USA) were retrieved from the Global Initiative on Sharing All Influenza Data (GISAID) database. A comprehensive dataset comprising a total of 298 SARS-CoV-2 sequences was assembled for subsequent analysis.

Sequence Analysis: Bioinformatics tools were employed for comprehensive sequence analysis, encompassing sequence alignment, phylogenetic tree construction, and mutation analysis. From those softawares we manged to work with MEGA 7, a desktop application for molecular evolutionary genetics analysis, facilitates the analysis of homologous gene sequences from multigene families or different species, focusing on inferring evolutionary relationships and DNA/protein evolution models [14]. Gblocks, a computer program, selectively removes poorly aligned positions and divergent regions from DNA or protein sequence alignments, enhancing the alignment quality for subsequent phylogenetic analysis. It follows reproducible conditions to select blocks based on conservation and gap density criteria, facilitating automation and reproducibility of phylogenetic analyses [15]. Additionally, GENIEGEN software allows for the analysis of DNA, RNA, and protein sequences, aiding in the discovery of genetic information expression, genotype-phenotype relationships, gene polymorphism, multigene families, and predictions in human genetics. It functions as a database of nucleic and peptide sequences, with the capability to incorporate new sequences [16]. Sequences were aligned using state-of-the-art alignment algorithms to ensure accurate alignment across the dataset. The resulting alignment served as the foundation for subsequent phylogenetic analyses.

Phylogenetic Analysis: Phylogenetic trees were constructed using robust methodologies to elucidate the evolutionary relationships among SARS-CoV-2 strains. Phylogenetic tree construction involved iterative processes, with sequences grouped based on the geographical location and data size of the countries under study to facilitate comparative analysis: Group 1: USA; Group 2: UK (United Kingdom); Group 3: AUKA (Australia, Kuwait, Saudi Arabia, Algeria); Group 4: GISA (Germany, Italy, Switzerland); Group 5: FESP (France, Spain). The construction of phylogenetic trees aimed to delineate the evolutionary dynamics of SARS-CoV-2 strains, providing insights into their geographical distribution and evolutionary origins.

Mutation Analysis: Mutations within SARS-CoV-2 genomes were systematically identified and analyzed to assess their potential impact on viral virulence and transmission dynamics. Comparative analysis of mutations between different clades enabled the identification of key genetic determinants associated with disease severity and transmissibility. Mutational landscape analysis provided critical insights into the evolutionary trajectory of SARS-CoV-2 and its adaptive mechanisms in response to selective pressures.

Statistical Analysis: Statistical methodologies were employed to quantify the significance of observed mutations and to assess their potential association with clinical outcomes. Comparative analyses between different clades and geographical regions were conducted to identify statistically significant differences in mutation frequencies and distributions.

Ethical Considerations: This study adhered to ethical guidelines for the use of genomic data, ensuring compliance with data-sharing policies and privacy regulations. All genomic data were anonymized and obtained from publicly available databases, with no identifiable information included in the analysis.

Results and Discussion

In our study, we focused on the phylogenetic analysis and comparison of COVID-19, which has garnered significant media coverage since its emergence in December 2019. Numerous laboratories have dedicated considerable time and effort to characterizing the virus using multiple genome-based techniques, aiming for a comprehensive understanding of the SARS-CoV-2 genome [17-20]. Nearly 300 sequences from various countries across different continents, obtained from the public GISAID database [21], were analyzed to achieve a clear resolution of the virus's diversity, evolution, mutations, and their positions within its genome.

Phylogenetic Analysis

Phylogenetic analysis of SARS-CoV-2 genomes unveiled a landscape rich in genetic diversity, reflecting the complex evolutionary dynamics of the virus. Within each country, distinct subclass groupings were discerned, underscoring the diverse evolutionary trajectories of SARS-CoV-2 strains across different

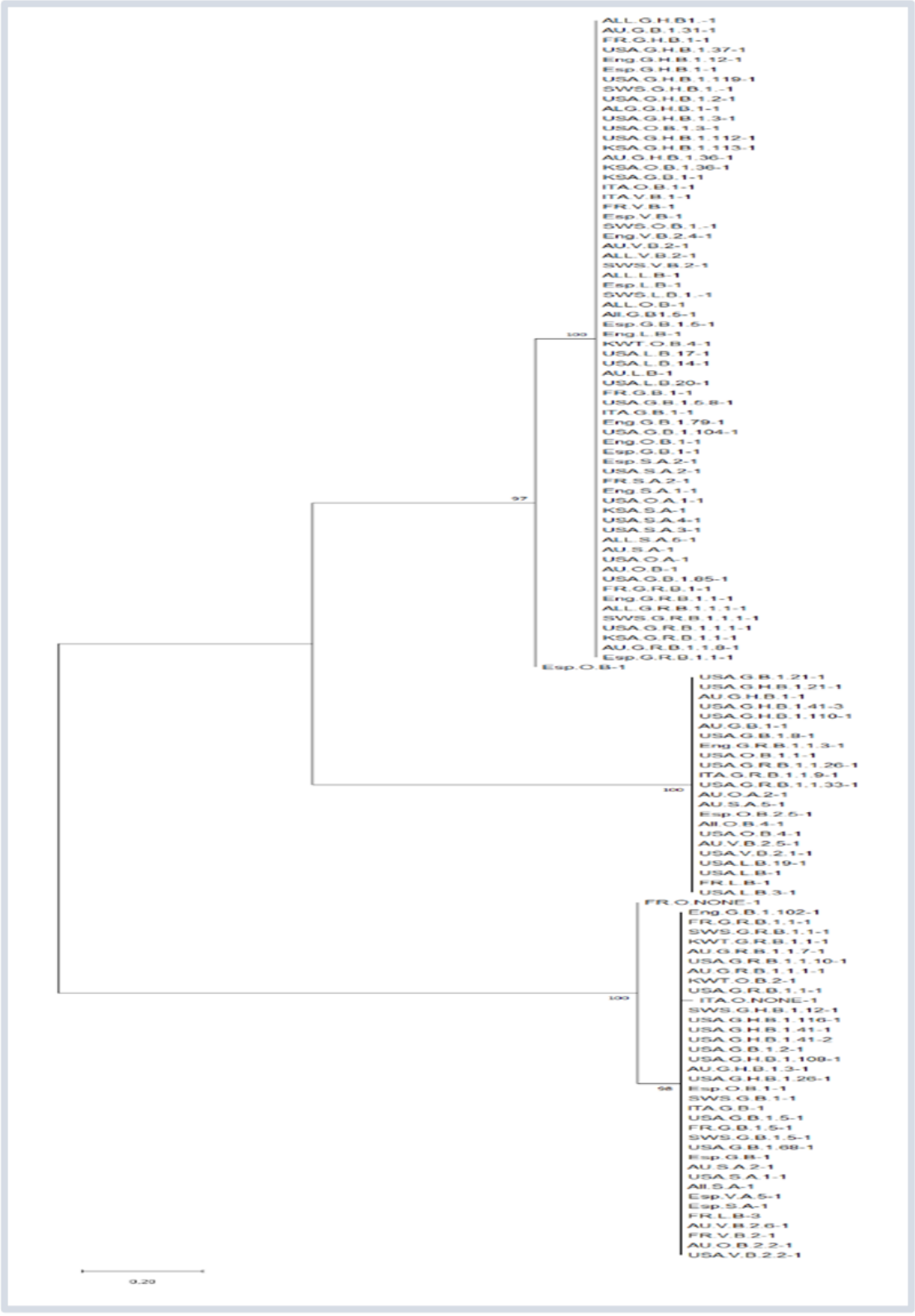

Figure 1: The phylogenetic tree grouping all countries of this study.

geographical regions. This intricate phylogenetic architecture highlights the dynamic nature of viral evolution and the potential for localized viral adaptation in response to environmental and host factors. Delving deeper into the phylogenetic trees, distinct patterns emerge within each country or region studied. For instance, in the USA (Group 1), a diverse array of viral lineages, encompassing clades G, GH, GR, S, L, V, and O, was observed, reflecting the complex epidemiological landscape of the pandemic. Similarly, phylogenetic analyses for other groups (ANG, AU, FE, ALL) unveiled a mosaic of viral lineages, underscoring the genetic heterogeneity inherent to SARS-CoV-2. The final phylogenetic tree (Figure 1) encompasses the results of previous steps and all the countries studied. It is divided into two classes, each further divided into subclasses grouping the different lineage sequences from the countries. We observed in the phylogenetic trees of this study that at the level of each subclass, there is no clear association with the region, lineage, or clade, reflecting the immense diversity and high mutability of this virus.

Mutation Analysis

A manual comparison of SARS-CoV-2 sequences using GenieGen software revealed a myriad of mutations scattered across the viral genome, indicative of ongoing genetic diversification. Notably, several mutations were identified that exhibited a significant association with increased virulence, implicating them as potential determinants of disease severity. Of particular interest were mutations observed in clade G and clade O sequences, which displayed distinct mutation profiles suggestive of differential virulence and transmissibility. These findings underscore the complex interplay between viral genetic variation and disease pathogenesis, highlighting the need for continued surveillance and monitoring of SARS-CoV-2 mutations for effective public health interventions. Table 1 presents the mutations identified through manual comparison of sequences from clades G and O for Group I countries, where mutations are observed in both clade O and clade G, as well as some mutations common to both clades, affecting various key viral proteins, such as ORF1ab, Spike (S) protein, and nucleocapsid (N) protein, suggests their potential role in shaping viral fitness and host interactions.

Table 2 shows the mutations determined from the manual comparison of sequences from clades G and O for Group II countries, where some mutations are exclusive to clade O, others to clade G, and some common to both. Additionally, there are a few mutations shared between Groups 1 and 2. Among the comparison results, 283 mutation positions are identified in the two preceding tables, along with 926 positions of rare mutations, indicating the high mutation rate this virus can undergo. Based on these findings, we suggest that mutations present in clade O may lead to a decrease in virus virulence. The differences in results between the two groups of countries allow us to conclude the significant diversity of this virus. Furthermore, comparative analyses between clades G and O shed light on specific mutations associated with each clade, hinting at potential differences in virulence and transmissibility. Noteworthy mutations identified in Group I and Group II countries, spanning crucial viral proteins, offer tantalizing insights into the evolutionary forces driving the diversification of SARS-CoV-2.

Implications for Public Health

The dynamic nature of SARS-CoV-2 evolution underscores the importance of vigilant surveillance and monitoring of genetic changes for the development of targeted public health interventions. By elucidating the genetic determinants of viral virulence and transmissibility, we can better inform the design of therapeutic strategies and vaccine development efforts. Moreover, the identification of specific mutations associated with increased mortality rates provides valuable insights into potential targets for therapeutic intervention, paving the way for the development of precision medicine approaches tailored to individual patient needs. Importantly, our findings corroborate previous studies [22, 23] suggesting differential virulence between clades G and O. The identification of specific mutations linked to virulence underscores the urgent need for targeted therapeutic interventions. Molecular docking studies targeting key mutations, particularly those associated with heightened virulence, hold promise as a strategy to mitigate the impact of the COVID-19 pandemic.

The study has several limitations that warrant consideration. Firstly, reliance on genomic data sourced from public databases introduces the potential for sampling bias, as certain geographic regions or demographic groups may be overrepresented or underrepresented. Additionally, variability in the accuracy and completeness of genomic data, along with potential sequencing errors or artifacts, could impact the reliability of mutation calls and phylogenetic reconstructions. While the study provides a snapshot of SARS-CoV-2 evolution at a specific time, ongoing viral evolution may lead to changes in genetic diversity and evolutionary relationships over time. Furthermore, establishing causal relationships between genetic variation and clinical outcomes requires additional experimental validation and clinical correlation studies, highlighting the need for caution in interpreting associations between mutations and virulence. Methodological constraints, such as algorithmic biases or assumptions, may also affect the robustness of the findings, necessitating careful validation and sensitivity analyses. Moreover, the generalizability of the findings may be limited to the specific dataset and analytical methods used, requiring replication in independent datasets and diverse populations for validation. Ethical considerations regarding data privacy, informed consent, and data sharing must also be addressed to safeguard individual rights and privacy. Finally, interpretation of phylogenetic trees and mutation profiles may be subject to bias, highlighting the importance of transparency and rigor in reporting methodologies and results.

Conclusions

In summary, this study offers a comprehensive examination of the genetic landscape, geographical distribution, and virulence characteristics of SARS-CoV-2 strains across a diverse array of countries. The intricate phylogenetic patterns observed underscore the dynamic nature of viral evolution and the capacity for adaptation to various environmental pressures. The results of this study

Table 1: Different mutations to identify from the comparison of the sequences of the countries of the Group 1.

Position	Clade G	Clade O	Mutation type	Repetition number	Region
8788	+	+	Substitution T/C	8 (O) / 2 (G)	ORF1ab
14811	-	+	Substitution T/C	10	ORF1ab
11089	-	+	Substitution T/G	9	ORF1ab
14414	+	+	Substitution C/T	10 (O) / 3 (G)	ORF1ab
20274	+	+	Substitution G/A	3 (O) / 9 (G)	ORF1ab
23394	-	+	Substitution A/G	4	Protein S
23403	+	+	Substitution G/A	15 (O) / 4 (G)	Protein S
23409	+	+	Substitution A/G	11 (O) / 1 (G)	Protein S
24868	+	-	Substitution G/A	4	Protein S
26150	-	+	Substitution T/G	8	ORF3a
25435	+	-	Substitution T/G	4	ORF3a
25569	+	+	Substitution T/G	4 (O) / 3 (G)	ORF3a
26720	-	+	Substitution C/G	2	Protein M
26536	+	-	Substitution C/A	2	Protein M
28083	-	+	Substitution C/G	2	ORF8
28150	-	+	Substitution C/T	6	ORF8
28151	-	+	Substitution C/T	2	ORF8
28317	-	+	Substitution C/T	5	Protein N
28694	-	+	Substitution C/T	3	Protein N
28688	-	+	Substitution C/T	3	Protein N
28881	+	+	Substitution A/G	6 (O) / 2 (G)	Protein N
28882	+	+	Substitution A/G	5 (O) / 2 (G)	Protein N
28883	+	+	Substitution C/G	5 (O) / 2 (G)	Protein N
28887	+	+	Substitution A/G	1 (O) / 3 (G)	Protein N
28888	+	+	Substitution A/G	1 (O) / 3 (G)	Protein N
28889	+	+	Substitution C/G	1 (O) / 3 (G)	Protein N

Table 2: The different mutations to be identified from the comparison of the sequences of the countries of Group 2.

Position	Clade G	Clade O	Mutation type	Repetition number	Region
3028	-	+	Substitution C/T	6	ORF1ab
3031	-	+	Substitution T/C	7	ORF1ab
3037	+	+	Substitution C/T	17 (O) / 4 (G)	ORF1ab
3336	-	+	Substitution C/T	8	ORF1ab
11083	-	+	Substitution T/G	11	ORF1ab
14805	-	+	Substitution T/C	4	ORF1ab
14408	+	+	Substitution C/T	10 (O) / 6 (G)	ORF1ab
23380	-	+	Substitution A/G	4	Protein S
23394	-	+	Substitution A/G	4	Protein S
23403	+	-	Substitution G/A	16 (O) / 4 (G)	Protein S
26720	-	+	Substitution C/G	2	Protein M
28144	-	+	Substitution C/T	3	ORF8
28688	-	+	Substitution C/T	3	Protein N
28881	+	+	Substitution A/G	6 (O) / 2 (G)	Protein N
28882	+	+	Substitution A/G	5 (O) / 2 (G)	Protein N
28883	+	+	Substitution C/G	5 (O) / 2 (G)	Protein N

underscore the dynamic nature of SARS-CoV-2 evolution and its implications for public health. Phylogenetic analysis revealed significant genetic diversity among viral strains, with distinct subclass groupings observed within each country, our analysis led us to conclude that globally, the virus's distribution does not correlate with regions, lineages, or clades at the subclass level, underscoring the virus's immense diversity and mutability. Moreover, the high mutation rate and mutations present in clade O suggest a potential cause for the virus's reduced virulence.

By elucidating the genetic determinants of viral virulence, this study provides crucial insights that can inform the development of targeted therapeutic interventions and vaccine strategies aimed at combating COVID-19. Furthermore, the identification of specific mutations offers promising avenues for further investigation through molecular docking studies, which may unveil potential therapeutic targets for drug development. Moving forward, sustained surveillance efforts are imperative to monitor the ongoing evolution and transmission dynamics of SARS-CoV-2, facilitating timely interventions and control measures to curb the spread of the pandemic and minimize its impact on global health.

References

[1] Atzrodt CL, Maknojia I, McCarthy RD, Oldfield TM, Po J, Ta KT, et al. A Guide to COVID-19: A global pandemic caused by the novel coronavirus SARS-CoV-2. FEBS J. 2020;287(17):3633-50. 10.1111/febs.15375

[2] Zatla I, Boublenza L, Hassaine H. SARS-CoV-2 morphology, genome, life cycle and our bodies' immune response: a review. Curr Top Virol. 2021. 18: 15-24 10.31300/CTVR.18.2021.15-24

[3] Ali SA, Baloch M, Ahmed N, Ali AA, Iqbal A. The outbreak of coronavirus disease 2019 (COVID-19)-An emerging global health threat. J Infect Public Health. 2020;13(4):644-6. 10.1016/j.jiph.2020.02.033

[4] Chatterjee, P., Nagi, N., Agarwal, A., Das, B., Banerjee, S., Sarkar, S., ... & Gangakhedkar, R. R. The 2019 novel coronavirus disease (COVID-19) pandemic: A review of the current evidence. The Indian journal of medical research, 2020. 151(2-3), 147.

[5] Valencia DN. Brief review on COVID-19: The 2020 pandemic caused by SARS-CoV-2. Cureus. 2020;12(3):e7386. 10.7759/cureus.7386

[6] El Zowalaty, M.E. and Järhult, J.D.. From SARS to COVID-19: A previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans–Call for a One Health approach. One health, 2020. 9: 100124.

[7] Zatla I, Boublenza L, Hassaine H. SARS-CoV-2 origin, classification and transmission: a mini-review. Curr Top Virol. 2021. 18: 31-38. 10.31300/CTVR.18.2021.31-38

[8] Alanagreh L.A., Alzoughool F. and Atoum M.. The human coronavirus disease COVID-19: its origin, characteristics, and insights into potential drugs and its mechanisms. Pathogens, 2020; 9(5), 331.

[9] Bonny, V., Maillard, A., Mousseaux, C., Plaçais, L., & Richier, Q. COVID-19: Pathophysiology of a Disease with Several Faces. The Internal Medicine Review. 2020.

[10] Saglietto, A., D'Ascenzo, F., Zoccai, G. B., & De Ferrari, G. M. COVID-19 in Europe: the Italian lesson. Lancet, 2020; 395 (10230): 1110-1111.

[11] Zatla I, Boublenza L, Boublenza A. Tracing the Origin and Early Progression of COVID-19 in Europe : An Epidemiological Descriptive Study. IEM 2023; 9 (3) :249-256. 10.61186/iem.9.3.249

[12] Zatla I., Boublenza L. The COVID-19 outbreak inception in Algeria: What Happened?. Algerian Journal of Health Sciences. 2023;18-15-24

[13] Martinez, I. L., Llinás, D. T., Romero, M. B., & Salazar, L. M. High Mutation Rate in SARS-CoV-2: Will It Hit Us the Same Way Forever. J Infect Dis Epidemiol. 2020; 6:176.

[14] Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016; 33(7):1870-4. 10.1093/molbev/msw054.

[15] Talavera, G., & Castresana, J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Systematic Biology. 2007.

[16] Marzin-Janvier, P. Supporting Experimental Design with Computer Environments: Studies in Genetics. RDST. Research in Didactics of Sciences and Technologies, 2015; 12: 87-112.

[17] Jaimes, J. A., André, N. M., Chappie, J. S., Millet, J. K., & Whittaker, G. R. Phylogenetic analysis and structural modeling of SARS-CoV-2 spike protein reveals an evolutionary distinct and proteolytically-sensitive activation loop. Journal of molecular biology. 2020.

[18] Suzuki, Y. Methods for Making Multiple Alignments of Genomic Sequences for Severe Acute Respiratory Syndrome Coronavirus 2. Meta Gene. 2020.

[19] Thieffry, D., & De Jong, H. Modeling, Analysis, and Simulation of Genetic Networks - Bioinformatics. Medicine/Sciences. 2002.

[20] Elayadeth-Meethal, M., Ollakkott S., Girish Varma G. COVID-19 and SARS-CoV-2: Molecular Genetics Perspectives. 2020.

[21] GISAID. Genomic epidemiology of novel coronavirus. 2020. Available at https://nextstrain.org/ncov

[22] Dao, T. L., Hoang, V. T., Colson, P., Lagier, J. C., Million, M., Raoult, D., Levasseur, A., & Gautret, P. SARS-CoV-2 Infectivity and Severity of COVID-19 According to SARS-CoV-2 Variants: Current Evidence. Journal of Clinical Medicine, 2021 10(12), 2635. 10.3390/jcm10122635

[23] Hamed, S. M., Elkhatib, W. F., Khairalla, A. S., & Noreddin, A. M. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. Scientific Reports. 2021;11(1), 1-8. 10.1038/s41598-021-87713-x

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ARTICLE | DOI: 10.5584/jiomics.v14i1.229

DTT protein equalization and Tryptophan protein quantification as a powerful tool in analytical proteomics.

Inês F. Domingos[1,2], Luís B. Carvalho[1,2], José L. Capelo[1,2], Carlos Lodeiro[1,2], Hugo M. Santos[1,2]*

[1]BIOSCOPE Research Group, LAQV-REQUIMTE, Department of Chemistry, [2]NOVA School of Science and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal, PROTEOMASS Scientific Society, 2825-466 Costa da Caparica, Portugal.

Available Online: 15 April 2024

Abstract

Assessing total protein levels in biological samples is a common procedure in biochemistry and molecular biology. In this study, we compare tryptophan fluorescence (WF) with Bradford and BCA assays to determine total protein in serum samples. Our results indicate that tryptophan fluorescence spectrometry is an efficient, sensitive, and straightforward technique for quantifying proteins in serum. We observed minimal variation between the three methods: BCA de one with the lowers LOD and LOQ. The tryptophan method offers the possibility of reusing the intact sample that does not need colourimetric reagents for quantification. Consequently, free tryptophan serves as a reliable universal standard. This assay can be performed using a conventional fluorescence spectrometer with cuvettes or in a 96-well plate format with a plate reader. The method was successfully used as proof of concept, using serum from patients diagnosed with myeloma and serum from healthy donors.

Keywords: DTT equalization, tryptophan, proteomics, serum, multiple myeloma.

Introduction

Quantifying total protein in biological samples illuminates interactions, signalling pathways, and cellular processes, aiding in diagnostics by comparing health versus disease conditions. Selecting a protein assay method is a complex decision, requiring consideration of the advantages and disadvantages inherent to each method, such as interferences, accuracy, reproducibility, sample handling and throughput. To minimize sample manipulation, researchers often maintain different protein assay methods in their laboratories to address diverse experimental requirements efficiently. In proteomics, protein determination is the prerequisite for optimal protein digestion and, subsequently, peptide characterization via mass spectrometry analysis.[1]

The Bradford protein assay was developed by Marion M. Bradford in 1976 [2]. It is a well-known colourimetric protein assay that relies on the change in absorption of the Coomassie Brilliant Blue G-250 dye. The Coomassie Brilliant Blue G-250 dye, existing in anionic, neutral, and cationic forms, undergoes a colour shift from red to blue under acidic conditions, binding to proteins during testing. The dye forms a strong non-covalent complex with the protein's carboxyl and amino groups through van der Waals forces and electrostatic interactions, exposing hydrophobic pockets in the protein's tertiary structure. The bound anionic form of the dye, maintained by hydrophobic and ionic interactions, exhibits a maximum absorption spectrum at 595 nm. The increase in absorbance at this wavelength is proportional to the quantity of dye that binds to amino acids, which can be linked to the protein concentration in the sample [3]. However, notable interferences include high concentrations of detergents, such as sodium dodecyl sulfate (SDS), commonly found in protein extracts used for cell lysis and protein denaturation. Moreover, the method is sensitive to time variations in sample incubation. In addition, the Coomassie Blue G250 dye binds preferentially to arginine and lysine protein groups, which may result in a varied assay response for different proteins. Despite these drawbacks, the Bradford assay remains widely used.

***Corresponding author:** Hugo M. Santos, email: hmsantos@fct.unl.pt

The Bicinchoninic Acid-based method for protein quantification, known as BCA method, was first published by Smith et al. in 1985. Later, in 2010, the Thermo Scientific™ Pierce™ BCA Protein Assay Kit [4] was introduced, which employs a detergent-compatible formulation based on bicinchoninic acid (BCA) for colourimetric protein detection and quantitation. This method uses the biuret reaction, in which proteins in an alkaline medium reduce Cu^{+2} to Cu^{+1}, which react with bicinchoninic acid, forming a purple-colored complex, which exhibits strong absorbance at 562 nm. The color formation in the BCA assay is influenced by the macromolecular structure of proteins, the number of peptide bonds, and the presence of specific amino acids (cysteine, tryptophan, and tyrosine). Limitations include incompatibility with reducing agents and metal chelators, although trace quantities may be tolerated, and reported responsiveness to common membrane lipids and phospholipids.

Tryptophan [5] exhibits intrinsic fluorescence due to its indole ring. This fluorescence arises from the absorption of ultraviolet (UV) light, typically around 280 nm, followed by the emission of light at longer wavelengths, generally around 340-350 nm. Once the quantum yield of tryptophan fluorescence is relatively high, makes it a sensitive probe for protein detection. Tryptophan fluorescence can be used for the quantification of proteins because the fluorescence intensity is proportional to the concentration of tryptophan-containing proteins. Furthermore, the requirement for lower sample volumes compared to alternative quantification methods makes tryptophan fluorescence assays suitable for situations where sample amount is limited. The assay is fast, contributing to streamlined laboratory workflows. Moreover, the method demonstrates minimal interference from substances commonly present in biological samples, ensuring more accurate results, especially in complex matrices.[5,6] Overall, the combination of real-time monitoring, low sample volume, quickness, and minimal interference positions tryptophan fluorescence assay as a valuable tool for protein quantification. In this work, we have used the three aforementioned methods to compare protein quantification in the serum of healthy individuals.

Materials and methods

Human serum samples: The serum samples from 11 healthy volunteers were used. Data about the study cohort are presented in Supplementary Material 1, Table 1.

Serum sample preparation: Serum samples were collected in red glass vacutainer tubes without anticoagulants or preserves. The samples were allowed to clot at room temperature, RT, followed by centrifugation at 2 000 x g for 10 minutes at RT. After centrifugation, serum was aliquoted and stored at -80 ºC.

Bradford assay: Bovine serum albumin (BSA) was used to generate a calibration curve ranging from 0.125 to 1.4 µg/µL. Briefly, BSA working solutions containing 25 to 280 µg were pipetted into 0.5 mL microtubes, followed by adding MQ water up to 200 µL. To quantify the serum samples, we initially prepared a 1:100 serum dilution in MQ-water to ensure that concentrations were within the linear range. Five µL of each standard and diluted serum were loaded in duplicates into each well, followed by the addition of 250 µL of Bradford Reagent. MQ-water was used as a blank. Afterwards, standards, samples and blanks were incubated at room temperature for 20 minutes. Finally, absorbances at 595 nm were measured using the Clariostar microplate reader.

Bicinchoninic acid assay (BCA): For protein quantification using the BCA assay, we employed the Pierce™ BCA Protein Assay Kit from Thermo Scientific with part number 23225. Following the manufacturer's protocol, we chose the microplate procedure as outlined below: 25 µL of each standard and samples were pipetted in duplicate into individual wells of a microplate (working range = 0.02 – 2 µg/µL). Subsequently, 200 µL of the Pierce reagent was added. The plate was gently mixed for 30 seconds, covered with aluminum foil, and then incubated for 30 minutes at 37 ºC. Finally, absorbances at 562 nm were measured using the Clariostar microplate reader.

Tryptophan emission assay: The standard calibration curve was created using 0.0102 µg/µL tryptophan dissolved in 8M Urea in 0.1M Tris-HCl pH 8 to span a linear detection range from 0 to 5.1x10-3 µg/µL. For measurement, 75 µL of calibration solutions were transferred to a quartz-bottom 96-well plate. A dilution protocol was applied to align with the assay's linear range to assess the concentration of proteome and proteome digest in unknown samples. This involved adding 5 µL of each sample in duplicate to the plate wells, followed by adding 70 µL of the 8M Urea in 0.1M Tris-HCl pH 8. Measurements were performed with an excitation wavelength set at 280 nm, with a long pass dichroic mirror at 309 nm and an emission wavelength of 350 nm, utilizing a bandwidth of 20 nm. The quantification of proteome and proteome digest was derived from fluorescence intensity measurements, which applied an average tryptophan weight content assumption of 1.17% for human proteins [5].

Depletion of abundant serum proteins: To 20 µL of the raw serum sample, 2.2 µL of 500 mM DTT was added, followed by incubation for 30 minutes at 37 ºC. This procedure was performed in triplicate. After incubation, samples were centrifuged at 20,000 x g for 20 minutes. Afterwards, the supernatants were withdrawn to new microtubes, and the pellets were gently washed with ten µL of MQ H2O. The washing factions were combined with the previous supernatants. Quantifying the protein concentration in the depleted samples: Supernatants (SN) were quantified via Tryptophan Emission as described previously. Quantifying the protein content in the pellet was done using the following formula: $\mu g\ of\ Protein\ in\ Pellet = \mu g\ of\ Protein\ in\ Raw\ Serum - \mu g\ of\ Protein\ in\ SN$ - Equation 1.

Protein reduction, alkylation and digestion: Sample preparation was performed as described previously, with optimization for the supernatant and pellet samples.[7,8] Initially, supernatant samples were diluted to achieve a target protein concentration of 103 ± 8 µg per 20 µL of sample volume. This step was followed by adding 5 µL of a Reduction/Alkylation solution consisting of 10 mM TCEP, 40 mM CAA, 0.1M Tris-HCl pH 8.8, to

each sample. The mixture was then incubated for 30 minutes at 37 °C. The pellets were solubilized with 150 µL of 70 mM TEAB followed by probe sonication for 1 minute (Ultrasonic frequency: 30kHz, Ultrasonic Amplitude: 100%, Cycle time: 0,8 s). Afterwards, 100 ± 2 µg of the pellets' proteins (final volume of 20 µL) were reduced with 5 µL of Reduction/Alkylation solution. Before trypsin digestion, samples were diluted to 150 µL with 70 mM TEAB.

For the proteome digestion phase, 5 µL of a Trypsin/Lysine-C solution, at a concentration of 0.67 µg/µL prepared in 70 mM TEAB, was added to each reduced and alkylated sample. The samples were left to digest overnight at 37 °C. Following digestion, the resultant peptide mixtures were concentrated by drying in a speed vacuum concentrator. Before downstream analysis, peptides were resolubilized in 150 µL of 3% (v/v) Acetonitrile (ACN) in 0.1% (v/v) aqueous formic acid (FAaq), followed by 10 minutes of sonication using an ultrasonic bath at 100% ultrasonic amplitude.

LC-MS/MS analysis was performed using UltiMate 3000 ultra-high performance liquid chromatographer from Thermo Scientific, coupled to Ultra High-Resolution Quadrupole Time-of-Flight (UHR-QTOF) IMPACT HD mass spectrometer from Bruker. 0.5 µL of the sample with a total peptide concentration of 0.6 µg/µL were loaded onto a µPAC™ Trapping column and desalted for 2.7 min with 1% (v/v) ACN in 0.1% FAaq at a flow rate of 15 µL min-1. Then the peptides were separated using an analytical column (200 cm µPACTM PharmaFluidics). with a linear gradient at 500 nL min-1 (mobile phase A: FAaq 0.1% (v/v); mobile phase B: 99.9% (v/v) ACN and 0.1% (v/v) FAaq) 0–2 min from 3% to 5% of mobile phase B, 5–76 min from 5% to 17% of mobile phase B, 76–104 17% to 25% B, 104–121 25% to 35% B. Chromatographic separation was carried out at 35 °C. MS acquisition was set to MS (2 Hz) cycles, followed by MS/MS (8–32Hz), cycle time 3.0 seconds, active exclusion, exclude after one spectrum, release after 2 min. The precursor was reconsidered if its current intensity was 3.0 higher than the previous intensity and intensity threshold for fragmentation of 2500 counts.

Bioinformatics data analysis and functional enrichment: Raw LC−MS/MS data were processed in MaxQuant (V.1.6.10.43) for protein identification and label-free quantification using standard settings.[9] Peptide lists were searched against the human Uniprot FASTA database. A contaminant database generated by the Andromeda search engine was configured with cysteine carbamidomethylation as a fixed modification and N terminal acetylation and methionine oxidation as variable modifications.[10] We set the false discovery rate (FDR) to 0.01 for protein and peptide levels with a minimum length of seven amino acids for peptides, and the FDR was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine as expected using trypsin. A maximum of 2 missed cleavages were allowed. Data processing was performed using Perseus (version 1.6.10.50) with default settings [11]. All proteins and peptides matching the reversed database were filtered out[12].

Results and Discussion

Comparison between Bradford-, BCA- and Tryptophan-based assays

To compare the Bradford, BCA, and Tryptophan assays, we used sera from 11 healthy individuals. For the Bradford and BCA assays, we used BSA to generate a calibration curve, as described in the material and methods section. For the tryptophan emission assay, the calibration curve was generated with L-Tryptophan. After the calibration curve was generated, we calculated the limit of detection (LOD) and limit of quantification (LOQ) for each assay. For the Bradford assay, the LOD was 0.002 µg/µL, and the LOQ 0.06 µg/µL. In contrast, for the BCA method, the LOD was 0.01 µg/µL, and the LOQ was 0.05 µg/µL. Finally, the tryptophan emission assay, with a LOD of 0.001 µg/µL and LOQ of 0.01 µg/µL, proved to be the most sensitive. Our LOQ for tryptophan-based assay matches those reported in the literature [5]. Subsequently, the protein content of the raw sera, pellets, and supernatants was determined in triplicate. Results are shown in Figure 1a and in Supplementary Material 1 Table 2. When comparing the assays, the Bradford provides less scattered data, while the BCA consistently provides slightly higher protein concentration values. The Bradford assay was found to be the most accurate. It is worth noting that the normal range for serum total protein concentration typically falls within the range of 60–80 g/L.[13] Figure 1a shows (i) that the Bradford assay analysis revealed a total protein concentration ranging between 57 and 70 g/L; (ii) that the BCA assay was found to present a more extensive range spanning from 62 to 81 g/L, and (iii) that the assessment conducted via the tryptophan emission assay exhibited a range of 56 to 76 g/L.

Assessment of the DTT-based serum protein equalization process

Serum constitutes a complex mixture of tens of thousands of proteins, some highly abundant while others are in significantly lower concentrations. For an in-depth analysis of plasma proteomics, it is imperative to employ a depletion or equalization strategy to target the less abundant, potentially more insightful, proteins. Treatment of serum with DTT results in a pellet (P) containing highly abundant proteins and a supernatant (SN) rich in less abundant protein, among them, and remarkably, immunoglobulins 14. We applied the DTT approach, tryptophan quantification and label-free quantitative proteomics to access the serum proteome of raw serum and the depletion faction (SN and P). The effect of DTT over raw serum is shown in Figures 1 b and 1c. For the proteomics analysis, we performed a comparison using Z score normalization of the log-2 transformed data derived from the mass spectrometry measurements. A multiple-sample ANOVA test was employed, utilizing a permutation-based false discovery rate (FDR) approach with a threshold of 0.05 FDR and S0 set to 0. The statistically significant different proteins were selected, and a Hierarchical cluster was generated using Euclidean Distance for both tree rows and columns as depicted in Figure 1 c, which shows in a heatmap the proteins differentially expressed among raw, the Ps and the SNs. The list of these proteins is depicted in Supplementary 1 Table 5. Thus, a group of proteins (in green) with lower levels than in the P or in the raw data can be seen in the SN. This group constitutes the most abundant proteins present in lower levels in the supernatant as they tend to

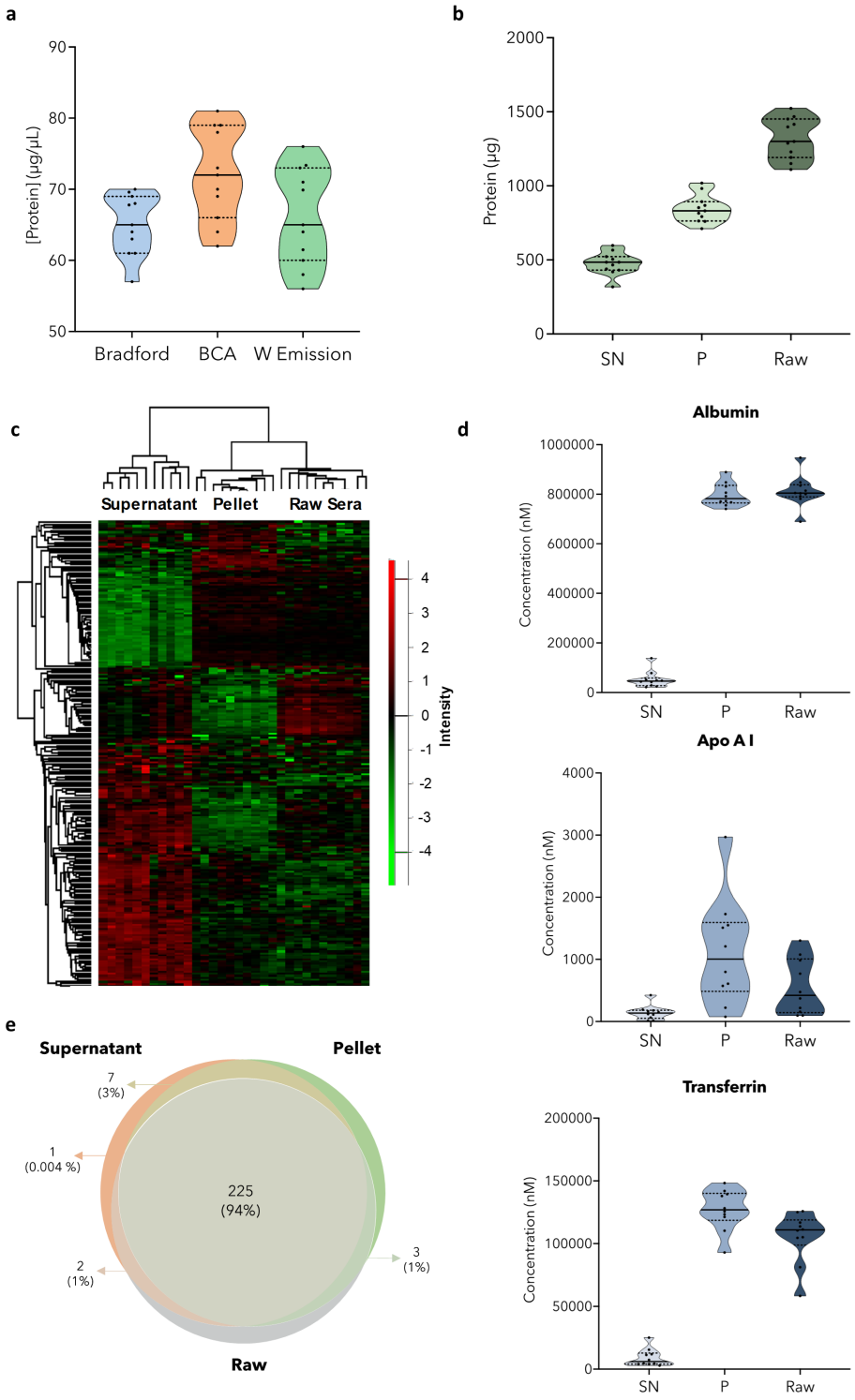

Figure 1 – Comprehensive proteomic analysis of human health serum. a) Violin plot of data distribution from three methods comparison in healthy individuals (n=11). Blue: Bradford assay. Orange: BCA assay. Green: tryptophan emission assay. Each sample was analyzed in duplicate. **b)** Distribution and density variation of protein mass (including two biological replicates) at each point for each fraction. The continuous bar in the middle represents the median. Discontinuous line represents the quartile lines. **c)** Hierarchical clustering of the three fractions analyzed (raw sera, supernatant and the pellet). Protein LFQ values were used to perform the cluster analysis (with average linkage, no constraint, preprocessing with k-means and Euclidean distance). **d)** Violin plot illustrating the data distribution of the three most abundant proteins in serum samples obtained from healthy individuals. SN corresponds to Supernatant, P corresponds to Pellet, and Raw corresponds to raw sera. Concentrations were derived from LC-MS/MS raw intensities using the total protein approach. **e)** Venn diagram illustrating the overlap of proteins identified in the raw serum, and in the depleted SN and Pellet fractions.

precipitate into the pellet. In this group of depleted proteins, we found, for instance, albumin, serotransferrin, a-1-Antitrypsin, plasminogen, complement-associated proteins, and apolipoproteins, which are some of the most abundant proteins present in serum. On the other hand, the less abundant proteins in the raw sera (in green) become more abundant in the supernatant (in red). This is a clear example of the highly abundant protein DTT equalization effect on the raw sera. This effect is further exemplified in the case of 3 individual proteins, as presented in Figure 1 d. Thus, albumin, the most abundant protein constituting circa 90% of the total protein content in serum, is found to be the most abundant protein in the pellet, with its level reduced about 16-fold in the supernatant. This is also an exciting characteristic of the DTT approach; instead of depleting the complete protein, some remains in solution. As albumin can also carry other proteins, its complete depletion from serum is not recommended. A similar situation can be observed for Apo A1 and Transferrin. The DTT works by equalizing the levels of the proteins in the supernatant rather than depleting a set of proteins. This concept is visualized in Figure 1e, where it can be seen that 94% of the proteins are commonly identified in raw serum, pellet, and supernatant.

Conclusions

The tryptophan assay is more sensitive than the Bradford and BCA analysis for proteomics purposes, with the advantage of simplicity and throughput. Furthermore, it is non-destructive, and the sample can be reutilized, further validating its use in proteomics. By applying DTT and tryptophan quantification, this study demonstrates the effectiveness of the DTT-based equalization process in revealing less abundant proteins, offering a promising avenue for deep proteomic analysis. The ability of DTT to maintain some proteins in solution, rather than completely depleting them, is particularly beneficial for preserving proteins like albumin that carry other proteins, thus avoiding their complete removal from serum analysis.

Acknowledgment

PROTEOMASS Scientific Society is acknowledged by the funding provided to the Laboratory for Biological Mass Spectrometry Isabel Moura (#PM001/2019 and #PM003/ 2016). This work received support from Fundação para a Ciência e a Tecnologia and Ministério da Ciência, Tecnologia e Ensino Superior (FCT/MCTES) through the projects LA/P/0008/2020 DOI 10.54499/LA/P/0008/2020, UIDP/50006/2020 DOI 10.54499/UIDP/50006/2020 and UIDB/50006/2020 DOI 10.54499/UIDB/50006/2020. H.M.S. acknowledges the Associate Laboratory for Green Chemistry-LAQV (LA/P/0008/2020) DOI 10.54499/LA/P/0008/2020 funded by FCT/MCTES for his research contract. L.B.C. is supported by the FCT/MCTES PhD grant 2019 (SFRH/BD/144222/2019).

References

[1] Krohn, R. I. The Colorimetric Detection and Quantitation of Total Protein. Curr Protoc Cell Biol 52, (2011).

[2] Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72, 248–254 (1976).

[3] Alexander J. Ninfa, David P. Ballou & Marilee Benore. Fundamental Laboratory Approaches for Biochemistry and Biotechnology. in 113 (2009).

[4] Smith, P. K. et al. Measurement of protein using bicinchoninic acid. Anal Biochem 150, 76–85 (1985).

[5] Wis□niewski, J. R. & Gaugaz, F. Z. Fast and sensitive total protein and peptide assays for proteomic analysis. Anal Chem 87, 4110–4116 (2015).

[6] Shelton, K. R. & Rogers, K. S. Tryptophanyl fluorescence of sodium dodecyl sulfate treated and 2-mercaptoethanol reduced proteins: A simple assay for tryptophan. Anal Biochem 44, 134–142 (1971).

[7] Fernández-Costa, C. et al. Sequential depletion of human serum for the search of osteoarthritis biomarkers. Proteome Sci 10, 55 (2012).

[8] Fernández, C., Santos, H. M., Ruíz-Romero, C., Blanco, F. J. & Capelo-Martínez, J. L. A comparison of depletion versus equalization for reducing high-abundance proteins in human serum. Electrophoresis 32, 2966–2974 (2011).

[9] Tyanova, S. et al. Visualization of LC-MS/MS proteomics data in MaxQuant. Proteomics 15, 1453–1456 (2015).

[10] Cox, J. et al. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J Proteome Res 10, 1794–1805 (2011).

[11] Tyanova, S. & Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. in 133–148 (2018). doi:10.1007/978-1-4939-7493-1_7.

[12] Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13, 731–740 (2016).

[13] Ministry of Health (Portugal). Intervalos de Referência (Convencional). https://www.acss.min-saude.pt/wp-content/uploads/2018/09/Tabela_Final.pdf (2018).

[14] Araújo, J. E. et al. Dithiothreitol-based protein equalization technology to unravel biomarkers for bladder cancer. Talanta 180, 36–46 (2018).