European Biophysics Journal
https://doi.org/10.1007/s00249-021-01528-2

BIOPHYSICS LETTER

Quality control of purified proteins to improve data quality and reproducibility: results from a large-scale survey

Nick Berrow[1] · Ario de Marco[2] · Mario Lebendiker[3] · Maria Garcia-Alai[4] · Stefan H. Knauer[5] · Blanca Lopez-Mendez[6] · André Matagne[7] · Annabel Parret[11] · Kim Remans[8] · Stephan Uebel[9] · Bertrand Raynal[10]

Received: 23 November 2020 / Revised: 23 February 2021 / Accepted: 26 March 2021

Abstract

As the scientific community strives to make published results more transparent and reliable, it has become obvious that poor data reproducibility can often be attributed to insufficient quality control of experimental reagents. In this context, proteins and peptides reagents require much stricter quality controls than those routinely performed on them in a significant proportion of research laboratories. Members of the ARBRE-MOBIEU and the P4EU networks have combined their expertise to generate guidelines for the evaluation of purified proteins used in life sciences and medical trials. These networks, representing more than 150 laboratories specialized in protein production and/or protein molecular biophysics, have implemented such guidelines in their respective laboratories. Over a one-year period, the network members evaluated the contribution these guidelines made toward obtaining more productive, robust and reproducible research by correlating the applied quality controls to given samples with the reliability and reproducibility of the scientific data obtained using these samples in follow-up experiments. The results indicate that QC guideline implementation facilitates the optimization of the protein purification process and improves the reliability of downstream experiments. It seems, therefore, that investing in protein QC might be advantageous to all the stakeholders in life sciences (researchers, editors, and funding agencies alike), because this practice improves data veracity and minimizes loss of valuable time and resources. In the light of these conclusions, the network members suggest that the implementation of these simple QC guidelines should become minimal reporting practice in the publication of data derived from the use of protein and peptide reagents.

Special Issue: COST Action CA15126, MOBIEU: Between atom and cell.

✉ Bertrand Raynal
bertrand.raynal@pasteur.fr

[1] Protein Expression Core Facility, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain

[2] Lab of Environmental and Life Sciences, University of Nova Gorica, 5271 Vipava, Vipava, Slovenia

[3] Protein Purification Facility, Wolfson Centre for Applied Structural Biology, Edmund J. Safra Campus—The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

[4] European Molecular Biology Laboratory (EMBL), Hamburg Outstation and Centre for Structural Systems Biology, 22607 Hamburg, Germany

[5] Biochemistry IV—Biopolymers, University of Bayreuth, 95440 Bayreuth, Germany

[6] Protein Production and Characterization Platform, Novo Nordisk Foundation Center for Protein Research, 2200 Copenhagen, Denmark

[7] Laboratory of Enzymology and Protein Folding, Centre for Protein Engineering, InBioS, University of Liège, Building B6C, Quartier Agora, Allée du 6 Août, 13, 4000 Liège, Belgium

[8] Protein Expression and Purification Core Facility, EMBL Heidelberg, 69117 Heidelberg, Germany

[9] Bioorganic Chemistry and Biophysics Core, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

[10] Plateforme de Biophysique Moléculaire, Department of Structural Biology and Chemistry, Institut Pasteur, 75015 Paris, France

[11] Present Address: Charles River Laboratories, Beerse, Belgium

Published online: 21 April 2021

Introduction

Research in the area of life science has been dramatically expanded, and technologically advanced, in recent years due to its increased relevance to sustaining developments in the fields of medicine. However, there are many examples of the inability to accurately reproduce published data (Begley and Ioannidis 2015; Marco et al. 2021; Freedman et al. 2015; Bradbury and Plückthun 2015; Baker 2016). This has been acknowledged by funding bodies (Collins and Tabak 2014) and journals (Announcement 2013, 2017; Free 2020), but, in contrast to other disciplines, in which reagents and methodologies have already been standardized, no specific standards for the quality controls of proteins used as reagents existed prior to this joint ARBRE-MOBIEU /P4EU initiative (Marco et al. 2021).

The advent of genome sequencing projects gave researchers worldwide access to the sequence data of their 'genes of interest' and the simplification in cloning and recombinant expression technology enabled the production of recombinant proteins for their own research. The downside of such a development is that many researchers, inexperienced in handling proteins, overlook some basic problems that can lead to flawed or artefactual observations when using proteins in experiments. Similarly, for the vast majority of commercially available protein reagents only minimal Quality Control (QC) data are normally provided, usually limited to protein sequence, Sodium Dodecyl Sulphate Poly-Acrylamide Gel Electrophoresis (SDS-PAGE) data and buffer composition. While this information is essential, it gives the end-user no guarantees that the protein will be soluble, correctly folded, functionally active, non-aggregated, etc. The deleterious effects of poor protein quality on data reproducibility are evident and protein quality control methods have been published previously (Lebendiker et al. 2014; Buckle 2011; Marco 2010; Raynal et al. 2014, 2021; Gräslund 2008; Medrano et al. 2012), but the research community has, thus far, not reacted to such endeavors. To this effect, the ARBRE-MOBIEU and the P4EU networks have produced a list of recommended tests (QC Guidelines) to aid in the validation of protein samples used in biological research. These Guidelines have been recently presented in detail (Marco et al. 2021) and are available at the webpages of both contributing networks (https://p4eu.org/protein-quality-standard-pqs and https://arbre-mobieu.eu/guidelines-on-protein-quality-control).

In this article, we assess, by means of a large-scale survey, how the implementation of a limited number of simple tests for the evaluation of the quality of protein and peptide reagents can significantly improve their quality, and consequently, the reproducibility and reliability of downstream experimental data derived from the use of such reagents. The results support our assertions (Marco xxxx) that data from these minimal QC tests should be presented in all publications to strengthen confidence in the protein/peptide reagents used in the experiments, in a very similar way to the statistical compliance currently already required by most scientific journals.

Methods

Guidelines as theoretical reference

A set of good practices and techniques suitable for protein QC and reported in the Guidelines (Marco et al. 2021) (https://p4eu.org/protein-quality-standard-pqs and https://arbre-mobieu.eu/guidelines-on-protein-quality-control) were assessed for their utility in daily laboratory practice, as they are aimed at determining protein sample quality (and therefore also the quality of the derived data) as well as data transparency. The proposed techniques were chosen because they are generally accessible and inexpensive. A reliable protein QC therefore requires the provision the following information:

(i) Identity of the protein of interest, full primary sequence and conditions used to produce and purify the protein.
(ii) Confirmation of protein purity, homogeneity, identity and integrity.
(iii) Complementary techniques are recommended for improved characterization, some of which should be considered mandatory for selected applications.

Survey

Members of both the ARBRE-MOBIEU and the P4EU networks participated in the collection of information on protein samples by adopting a common on-line survey sheet (Supplementary Material 1), in which the results of the QC tests, as applied to various samples, were recorded. The samples were proposed randomly by the participating laboratories to increase their heterogeneity. The number of samples for each single laboratory was limited to 5 to avoid possible bias due to the over-representation of the processes of few specific laboratories. In addition, the relative success of the downstream experiments performed using these samples was reported after self-evaluation by the participating laboratories. This second analysis step enabled the evaluation of the impact of the protein sample QC tests performed on the quality of the final experimental results obtained.

Statistics

For a chosen confidence value, the number of sample measures in each test determines the margin of error on each test. The margin of error E of the survey results was calculated according to the following formula for dichotomous outcome:

$$n = p(1-p)\left(\frac{Z}{E}\right)^2$$

where n is the number of samples, Z is the value from the standard normal distribution reflecting the confidence level that will be used (e.g., $Z = 1.96$ for 95%) and E is the margin of error. p is the proportion of successes in the population. All the calculations were done with a confidence level of 95% using a Z of 1.96.

To estimate the significance of the difference in the success of the downstream application, a two independent sample for dichotomous outcome statistical test was used. For each QC test and group of tests for which the downstream application was evaluated, the difference in succeeding in the downstream application between the failing and passed samples was statistically evaluated with the following formula

$$Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where n_1 and n_2 are the number of failing and passing samples in each QC test or group of tests, respectively. p_1 and p_2 are the proportion of samples being successful in their downstream application for the failing and passing sample, respectively. The null hypothesis of no difference was rejected for confidence higher than 95% i.e. $Z > 1.96$.

Results and discussion

Data analysis

Information on the test results of 188 samples, sourced from 43 academic laboratories worldwide (Supplementary Material 2), was collected and analyzed (see Supplementary Material 1 for details). Each laboratory entered 4–5 samples that were produced in either protein core facilities (23%) or academic research laboratories (77%). The academic research laboratory sample information was collected by the biophysical core facility of Institut Pasteur. According to statistical analysis, using dichotomous test this number of samples was sufficient to get an error margin up to 7% with a confidence of 95%. Concerning the source organisms of the proteins, one third of the tested proteins were bacterial, another third was of human origin, whereas mouse, plant and virus proteins represented each 6% of the submitted samples. The last 20% corresponded to proteins from a large array of different organisms, such as yeasts and archaebacteria, and some of them were even synthetic. The proteins were produced in different hosts: 78% in bacteria, 10% in mammalian, 8% in insect cells, and 4% were produced in yeast and cell-free systems. Proteins were used for different applications or for combinations of applications such as biochemical analyses (~ 62%), structural studies (~ 42%), as recombinant antibodies (~ 10%), or for use in in vivo and activity tests (~ 15%).

According to the results of the survey, more than 98% of the tested samples were well described in terms of origin, primary sequence composition, buffer composition and storage conditions. However, the cloning strategy was unknown in more than 18% of the cases. This condition made it impossible to verify whether the chosen strategy resulted in unwanted modifications of the primary sequence. Furthermore, it impeded the ability to reproduce the experiment and undermined the overall trustworthiness on the samples. This observation confirmed the necessity of providing such information when publishing as it is not currently reported systematically.

Only two thirds of the samples were tested for identity and integrity (Fig. 1a) by one of the methods reported in the Guidelines (mass spectrometry was used in 83% of the cases) and 20% of these had an unexpected mass or sequence despite the protein providers confirming (or claiming to know) the exact sequence of the expression clone (Fig. 1b). In our experience, more accurate scrutiny by the protein provider is usually needed. In several cases, the discrepancy derived from errors in the provided sequence (absence of the tag, linker, etc.) This meant that several providers did not know the exact design of their gene construct with the consequence that the corresponding proteins can produce misleading results in the downstream application. We assessed the statistic implication of this drawback for the downstream applications (Fig. 2a) and a significant difference (53%) was observed between failing and acceptable samples, with an improved chance of downstream success for those with confirmed identity. This indicates the necessity of performing these preliminary analyses, as it seems that the sample identity is too often a contributing factor to poor quality of protein reagents. MS is now readily accessible to most researchers, either through core facilities or commercial providers, and should be considered more widely as an essential standard, QC for proteins and peptide reagents.

The purity of the samples was tested in 97% of the cases (Fig. 1a) and most of the participants (93% of tested samples) relied on SDS-PAGE to assess this parameter, as it is a commonly available technique in almost any research laboratory. However, special care should be taken in the

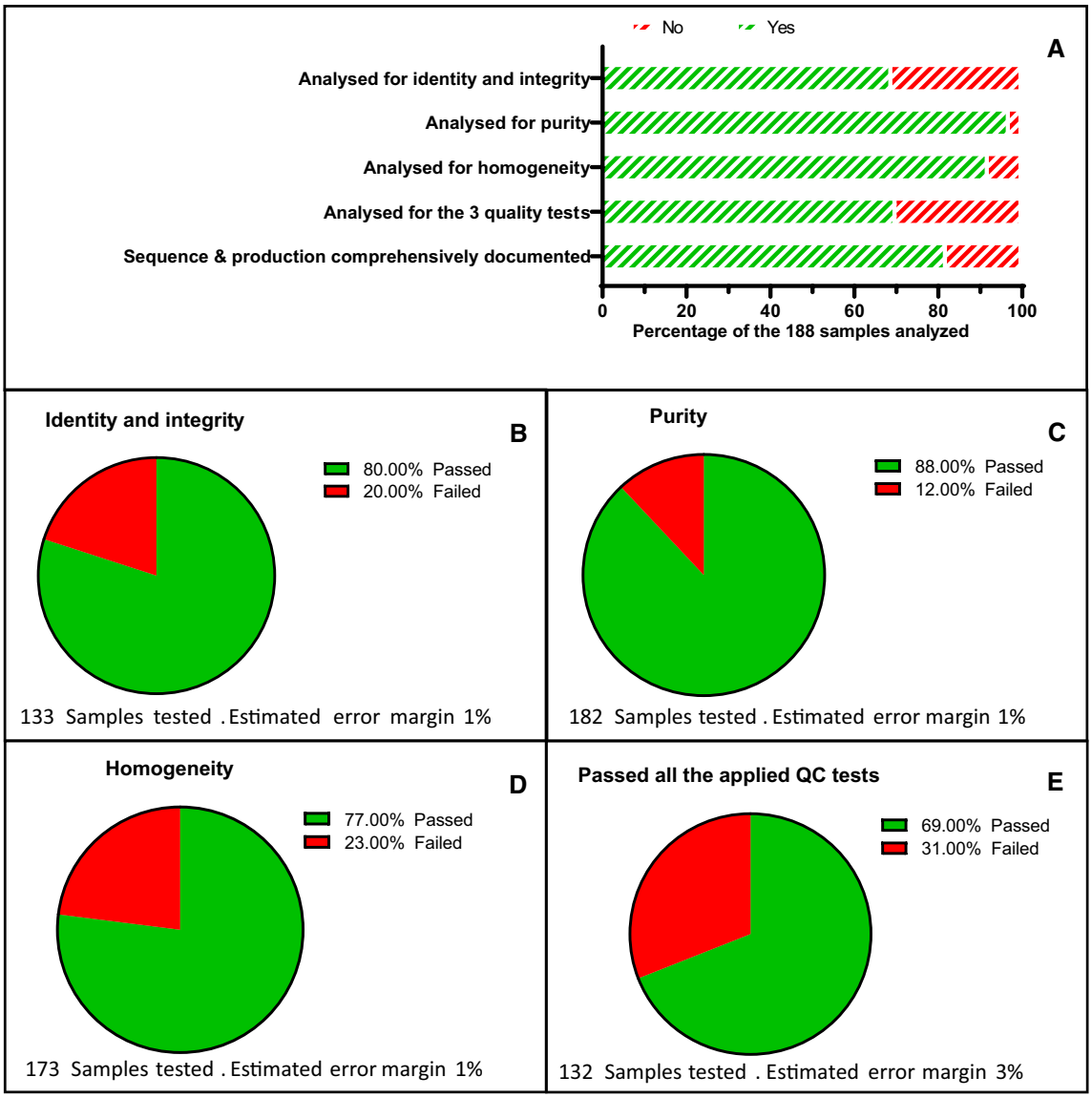

Fig. 1 Summary of the provided information and the tests performed with the samples. The error margin was estimated according to statistical formula for one sample with dichotomous outcome. **a** Information and analyses realized for the 188 samples. "Sequence & Production Comprehensively documented" means that protein samples are correctly described and with sufficient information to reproduce the sample. "Identity and Integrity"-samples were evaluated by MS, "Analyzed for Purity" refers to samples assessed by appropriate techniques such as SDS-PAGE, CE, RPLC or equivalent. "Analyzed for homogeneity"-samples were evaluated for homogeneity using at least one of the following techniques: SEC, DLS, SEC-MALS, Field-Flow Fractionation, Field-Flow Fractionation-MALS, Analytical Ultra-Centrifugation. **b–d** Breakdown of sample testing results. Percentage of the samples that failed/passed each of the three main quality testing categories criteria (identity and integrity, purity, homogeneity respectively). **e** Percentage of samples that satisfied tests in all three testing categories

choice of staining reagents used since their sensitivity varies from 1 to 100ng (Buckle 2011). Consequently, the possibility to detect minor impurities, representing from 1 to 5% of the total sample, is directly dependent on the staining method. According to the survey results, 12% of the samples tested by SDS-PAGE were evaluated as impure (Fig. 1c) as they contained proteins of unexpected molecular weight. Only 27% of these contaminated samples gave fully satisfactory results in their downstream application (Fig. 2b), namely reproducible data confirming those obtained with a positive control or expected results according to complementary analyses. In comparison, when samples classified as pure were used in downstream applications, 66% of the results were assessed as satisfactory. This strong correlation between sample purity and

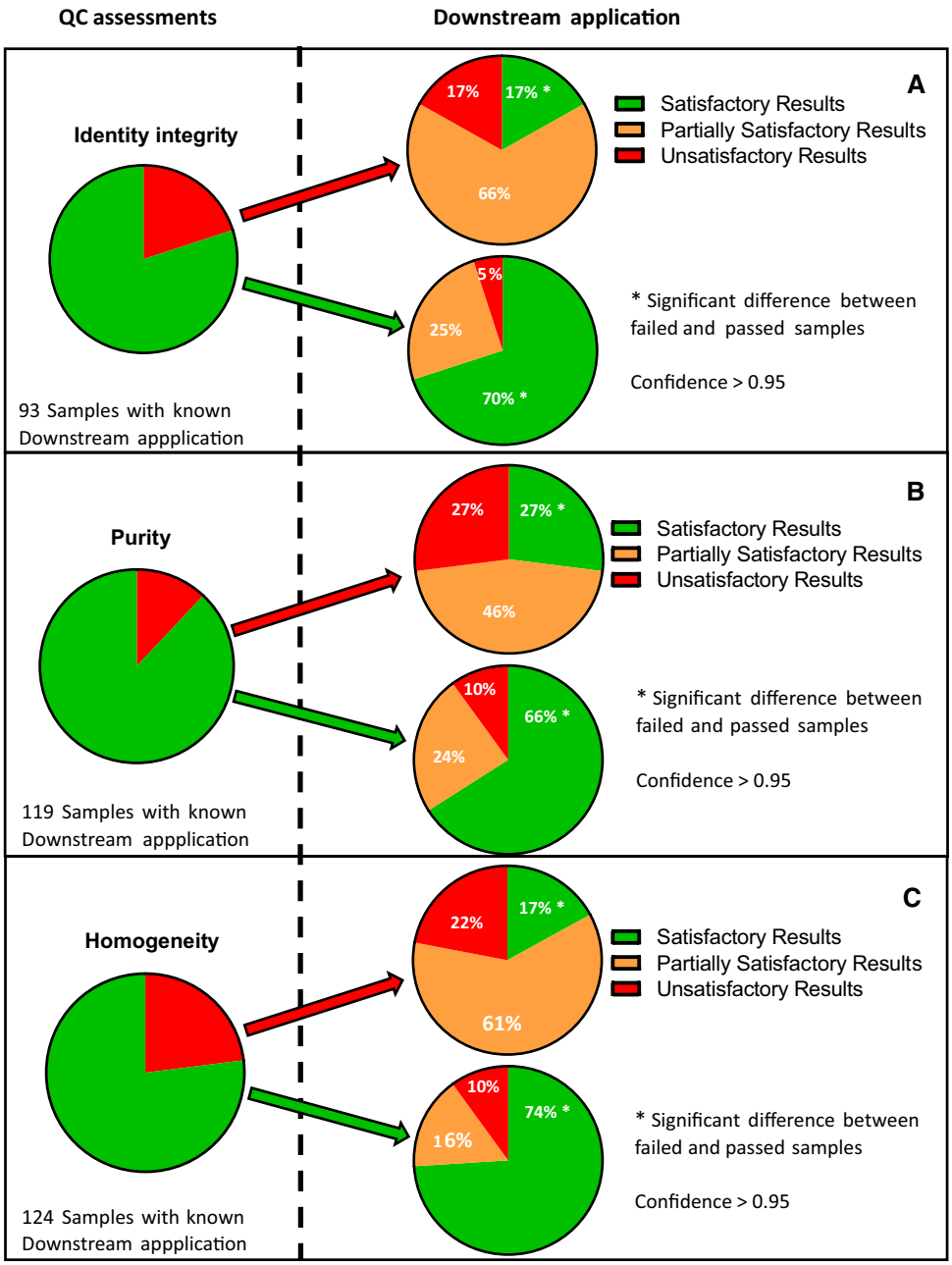

Fig. 2 Correlation between QC assessment and level of satisfactory results obtained in downstream applications. The assessments were presented in Fig. 1. Only the data for which the downstream application was known were used. The confidence was estimated according to statistical formula for two samples with dichotomous outcome

the success of the downstream experiments confirms the necessity of this quality evaluation step.

The homogeneity (i.e., size distribution) of the samples was tested primarily by dynamic light scattering (DLS) and/or analytical size exclusion chromatography (SEC) and in 20% of the tests performed the chromatography was coupled to a static light scattering detector (SEC-MALS). 8% of the samples were not assessed for homogeneity by any method (Fig. 1a), which is surprising as SEC is usually used by laboratories as a secondary or final purification step for protein production. The limited exploitation of these simple and generally accessible methodologies represents a striking case of missed opportunity since 23% of the tested protein samples were not homogeneous but presented several size-defined species and/or aggregates (Fig. 1d). We observed that the expression system had a strong influence on the homogeneity of the purified sample. 27% of the samples produced in bacteria were not homogeneous, in comparison to 15 and 10% of the proteins produced in mammalian and insect cells, respectively. Among the samples that were evaluated as homogeneous, 74% were assessed to have performed successfully in their downstream applications, whereas only 17% of the non-homogeneous samples performed successfully in their downstream applications

(Fig. 2c). These results show clearly that homogeneity is an important factor in determining the successful achievement of reproducible research. Furthermore, the detection of poor sample homogeneity indicated the necessity of improving this sample characteristic. During this study, the homogeneity of nine samples was substantially improved by optimizing the buffer composition. In terms of homogeneity, after the optimization of buffer composition, protein concentration and storage conditions using DLS and differential scanning fluorimetry (DSF) (Raynal et al. 2014; Medrano et al. 2012; Monsellier and Bedouelle 2005; Dupeux et al. 2011), the quality of 95% of these samples was improved sufficiently for them to be used in their downstream applications.

UV absorbance spectroscopy also proved effective in evaluating sample homogeneity and stability. Thus, the assessment of a sample by recording the 'full' UV spectrum (240–400 nm wavelengths), instead of a simple A_{280} nm reading to assess concentration, allowed the detection of aggregates in 75% of the non-homogenous samples, as well as the presence of nucleic acid and trace quantities of reducing agents in the buffer. This simple UV spectral scan, that is recommended as an extended technique and was performed on 60% of the samples, is rapid, preserves the sample and relies on spectrophotometers that are usually available in most of the laboratories. Among the samples that did not have a satisfactory UV spectrum, only 17% fully succeeded in their downstream applications. In contrast, 69% of the samples classified as non-aggregated by the UV spectrum provided convincing downstream data.

Overall, a third of the samples out of 132 failed at least one of the minimal tests (Fig. 1e), with 21% failing only one test, 6% two tests and 4% all the tests. In the absence of the minimal controls reported in the guidelines, researchers would have used preparations that were contaminated, degraded, or aggregated in approximately one third of their experiments. This practice would certainly have resulted in misleading or irreproducible results. The consequences of using protein samples that failed to pass at least one of the quality control tests in downstream applications are summarized in Fig. 3. Samples that failed one of the minimal tests performed successfully in only 21% of the downstream applications, whereas 79% were deemed to have produced inconsistent or partially satisfactory results. In contrast, the samples that passed all the minimal tests were deemed to have performed successfully in 74% of the downstream applications, whereas only 26% produced inconsistent or partially satisfactory results. This clearly shows that using these simple assessment and optimization procedures, the probability of the proteins performing successfully in the downstream procedure increases drastically.

The detailed evaluation of the data derived from the analysis of these samples is summarized below. However, we would like to point out that several participants to the survey work in specialized laboratories that are used to good practices for protein quality evaluation and already implemented several of the QC analyses reported in the Guidelines. We were therefore surprised that even in such a context, a systematic application of the required techniques could lead to the significant improvement of the sample quality we observed. Consequently, we expect that, when applied by labs that do not routinely perform QC tests, the introduction of the good practices indicated in the Guidelines will impact

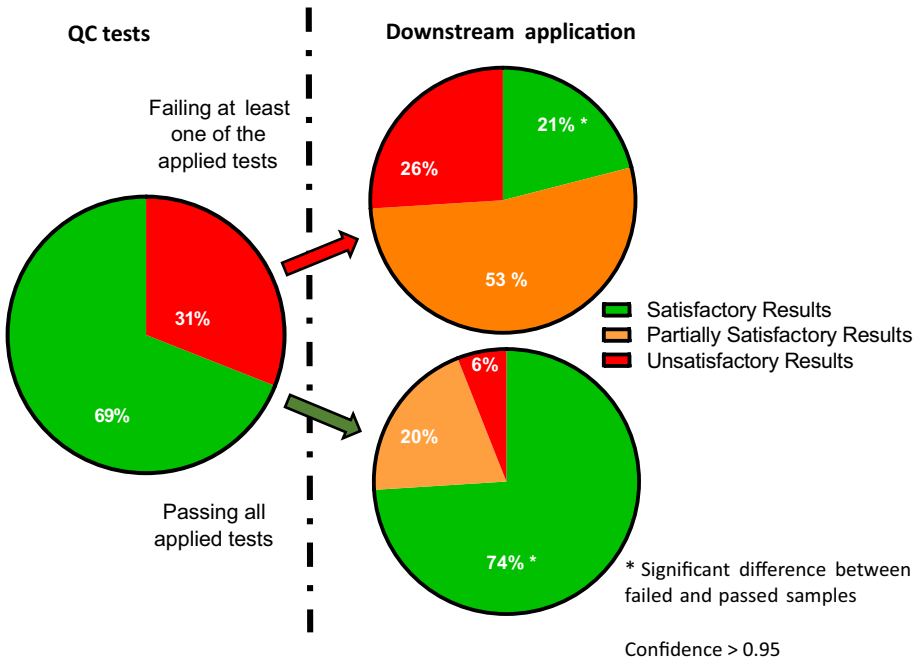

Fig. 3 Overall correlation between favorable QC results and level of satisfactory results obtained in downstream applications. Samples were categorized as either passed all or failed at least one of the applied minimal QC tests (Fig. 1, left chart). Sample users evaluated and reported the downstream results obtained with their protein samples choosing among three categories: Unsatisfactory, Ambiguous/Partially Satisfactory or Satisfactory. Finally, sample downstream reliability of sample performance in downstream applications was correlated to their QC output for 64 of them (fulfilling all the applied QC tests or failing one or more QC tests)

the quality of the produced proteins and the results obtained with these protein reagents even more strongly.

i. Confirmation of construct identity: More than 98% of the tested samples were well-described in terms of origin, primary structure, buffer composition, and storage conditions.
ii. Purity assessment: This simple assessment identified 12% of the tested samples as 'impure', as they contained significant levels of proteins of unexpected molecular masses.

Homogeneity/dispersity: The combined SEC/DLS evaluation of sample dispersity already revealed that nearly 25% of the tested protein samples were not homogeneous, but contained both low and higher-order aggregates. In the case of nine of these samples identified as 'non-homogenous', attempts were made to improve dispersity, which in all cases led to better homogeneity.

Identity and intactness: Only 66% of the samples were tested for identity and/or integrity (by MS and/or Edman sequencing) and, out of these, 20% gave masses or sequences that did not correspond to the anticipated values.

Conclusion

These data clearly demonstrate that the application of a limited number of simple QC tests provides reliable indicators of the quality of the protein reagents, with the more stringently assessed reagents producing significantly improved, more reproducible results in downstream applications. The quality control processes outlined here are simple, with instruments or services accessible to most laboratories, their costs are limited (less than 100€) and ultimately cost-effective when compared to the overall cost of the experiment and the risk of producing poor quality, irreproducible or artefactual data.

In addition, the data derived from such tests can be used to inform researchers on the nature of a problem, e.g. protein solubility or aggregation state, and allow the problem to be corrected, or improved upon, in a logical, informed manner (e.g. see refs (Free 2020; Marco et al. 2021; Lebendiker et al. 2014; Raynal et al. 2014; Raynal et al. 2021) for defined decision-making processes for improving protein quality).

The implementation of these simple QC steps, and the public availability of such QC data, could therefore significantly increase the level of confidence in the published data resulting from the use of these reagents. Currently, the published information on protein sample provenance and quality data often still falls short of a full description of the experimental conditions used and therefore the means for other researchers to reliably reproduce the experimental data are frequently lacking. With the advent of on-line publishing and the corresponding removal of space constraints on paper content, it is essential that these data are integrated into the Material and Methods or Supplementary Data sections when publishing as suggested in ref. 10. It is our assertion that these published sections should therefore contain sufficient information to faithfully reproduce the data therein, without resorting to making contact with the authors. It should also contain details of the QC tests performed on any protein/peptide reagents used in a study in order to give referees and readers an indication of the quality of the materials being used to derive any given data set.

It is evident from the presented data that a small investment of time in the QC of protein/peptide-related reagents, at negligible additional costs, can produce a significant improvement in the quality and reproducibility of experimental data obtained using these reagents. The feasibility of the simple tests suggested here leaves little room for excuses for 'non-compliance' with the minimal protein QC procedures (Free 2020) that should be considered as an essential component during the process of submitting a research paper for publication.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00249-021-01528-2.

Acknowledgements ARBRE-MOBIEU is supported by European CO-operation in Science and Technology (COST) Action number CA15126. We thank all the collaborating laboratories (see Supplementary Materials 2) for providing results on their samples.

References

Announcement (2013) Reducing our irreproducibility. Nature 496:398
Announcement (2017) Towards greater reproducibility for life-sciences research in Nature. Nature 546:8
Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454
Begley CG, Ioannidis JP (2015) Reproducibility in science: improving the standard for basic and preclinical research. Circ Res 116:116–126
Bradbury A, Plückthun A (2015) Reproducibility: standardize antibodies used in research. Nature 518:27–29
Buckle AM et al (2011) Recombinant protein quality evaluation: proposal for a minimal information standard. Stand Genomic Sci 5:195–197
Collins FS, Tabak LA (2014) NIH plans to enhance reproducibility. Nature 505:612–613
de Marco A (2010) Reagent validation: an underestimated issue in laboratory practice. J Mol Recognit 23:136
de Marco A, Berrow N, Lebendiker M, Garcia-Alai M, Knauer SH, Lopez-Mendez B, Matagne A, Parret A, Remans K, Uebel S, Raynal B (2021) Quality control of protein reagents for the improvement of research data reproducibility. Nat commun **(in press)**
Dupeux F, Röwer M, Seroul G, Blot D, Márquez JA (2011) A thermal stability assay can help to estimate the crystallization likelihood of biological samples. Acta Crystallogr D Biol Crystallogr 67:915

Free T (2020) The ABCs of reproducibility; effecting attitudes, behaviors and change. Biotechniques 69:359–361

Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. PLoS Biol 13:e1002165

Gräslund S (2008) Protein production and purification. Nat Methods 5:135–146

Lebendiker M, Danieli T, de Marco A (2014) The Trip Adviser guide to the protein science world: a proposal to improve the awareness concerning the quality of recombinant proteins. BMC Res Notes 7:585

Medrano G, Dolan MC, Condori J, Radin DN, Cramer CL (2012) Quality assessment of recombinant proteins produced in plants. In: Lrence A (ed) Recombinant gene expression: methods and applications. Methods Mol Biol 824:535–64

Monsellier E, Bedouelle H (2005) Quantitative measurement of protein stability from unfolding equilibria monitored with the fluorescence maximum wavelength. Protein Eng Des Sel 18:445

Raynal B, Lenormand P, Baron B, Hoos S, England P (2014) Quality assessment and optimization of purified protein samples: why and how? Microb Cell Fact 13:180

Raynal B, Brûlé S, Uebel S, Knauer SH (2021) Assessing and Improving protein sample quality. In: Daviter T, Williams M (eds) Protein-ligand interactions: methods and applications. Springer Science+Business Media, LLC (**in press**)

Nature collections: special 18 October 2018 Challenges in irreproducible research. https://www.nature.com/collections/prbfkwmwvz. Accessed 20 Apr 2021