Methodological insights on proteogenomic approaches to enhance proteomics

Laura Fancello[1] and Thomas Burger[1]

[1] *Univ. Grenoble Alpes, CNRS, CEA, Inserm, Profi FR2048, Grenoble, France.*

Abstract

Proteogenomics aims at identifying variant or unknown proteins in bottom-up proteomics, searching transcriptome- or genome-derived custom protein databases. However, empirical observations reported that the large size of these proteogenomic databases is associated to lower sensitivity of peptide identifications. Various strategies were proposed to avoid this, including the generation of reduced transcriptome-informed protein databases (*i.e.*, built from reference protein databases only retaining proteins with expressed transcript in the sample-matched transcriptome), which were found to increase peptide identification sensitivity. In this work, we propose a detailed evaluation of this approach. First, we establish that the increased sensitivity in peptide identification is in fact a statistical artefact, which directly results from the limited capability of TDC to accurately model incorrect target matches with excessively small databases. As anti-conservative FDRs likely hamper the robustness of the resulting biological conclusions, we advocate for alternative FDR control methods that are less sensitive to database size. Second, we show that despite not increasing sensitivity, reduced transcriptome-informed databases are useful, as they allow reducing ambiguity of protein identifications, yielding fewer shared peptides. Furthermore, we illustrate that searching the reference database and subsequently filtering proteins with unexpressed transcript similarly reduces protein

identification ambiguity, while representing a more transparent and reproducible strategy. To summarize, using reduced transcriptome-informed databases is an interesting strategy that has not been promoted for the good reason (an artifactual peptide identification sensitivity increment instead of a protein identification ambiguity decrement).

Keywords: proteogenomics, proteomics, transcriptome-informed protein databases, peptide identification sensitivity, protein identification ambiguity, FDR control, target-decoy competition

BACKGROUND

The term “proteogenomics” originally referred to the use of proteomics to enhance genome annotation, by inferring coding genomic regions based on evidence from mass spectrometry-based proteomics[1]. Nowadays, it more broadly indicates the combined analysis of genomics and/or transcriptomics together with proteomics in a large spectrum of applications: from the study of gene expression regulation at transcript and protein level, to the identification of specific protein variants expressed in cancer[2,3]. Most importantly, proteogenomics represents an attractive strategy to enhance proteomics in two main ways: *i.* improving protein inference; *ii.* improving database searches for peptide identification.

Protein inference is a central issue in proteomics, given the presence of shared peptides. This nominates peptides that might originate from different proteins sharing homology; or from different proteoforms due to alternative mRNA splicing, post-translational modifications, proteolytic cleavages, and/or allelic variants. Indeed, in bottom-up mass spectrometry-based proteomics, the most widely used proteomic approach,

peptide-protein connectivity is lost for experimental reasons and protein identifications are to be inferred from peptide identifications. Traditionally, the issue of protein inference was addressed using simple heuristics, such as the two-peptide rule (only proteins identified by at least two peptides are retained) or the parsimonious principle (the smallest subset of proteins which can explain most or all peptides is retained)[4,5]. Later on, more refined probability-based approaches were developed, which model shared peptide mappings to their parent proteins[6–9]. Most commonly, when proteins cannot be discriminated based on peptide identifications (*i.e.*, they are identified by the same set of peptides) they are reported as a protein group, which complicates comparisons between different experiments and protein quantification. In this context, proteogenomics can aid protein inference using evidence from transcript expression: in particular, some Bayesian approaches were developed based on this strategy[10–12]. The other main contribution of proteogenomics to proteomics relates to the refinement of reference protein databases used for peptide identification. Classically, peptides resulting from bottom-up analyses are identified by matching their experimentally measured mass spectra against theoretical spectra of all candidate peptides from a user-selected reference protein database. The underlying assumption is that such a database exhaustively and accurately describes all protein sequences present in the sample. However, this may be unrealistic for two reasons. First, reference databases only contain canonical -- experimentally validated or predicted -- protein sequences, while other variants or isoforms may be present, especially in tumour samples. Second, a reference protein database may simply be lacking for less studied organisms with scarce or no genomic annotations. In the first case, more exhaustive protein databases including undocumented or variant peptides can be generated by appending to the reference database variant sequences from public genomic repositories (*i.e.*, COSMIC or dbSNP)[13–15] or sample-specific variants identified from

matched transcriptomes or genomes. In the second case, protein databases can be generated by 6-frame translation of the genome or of the sample-matched transcriptome[16,17]. A major downfall of these proteogenomic databases is represented by their typically large size. Searching very large databases represents a considerable computational load and complicates the task of discriminating between correct and incorrect matches. In particular, various works showed that when using target-decoy competition (TDC) for FDR control on large database searches, fewer peptides are identified at the same FDR level, which stands in stark contrast with the initial motivation of proteogenomics[18–20]. To avoid this, it was proposed to perform separate FDR validation of canonical and novel peptides and to apply post-search filters or machine learning methods to increase confidence in the newly identified peptides[13–15,21,22]. Additionally, various strategies were adopted to limit the size of databases generated by proteogenomics. When possible, genome 6-frame translation was replaced by translating candidate ORFs identified by gene prediction algorithms; sample-specific variants from matched sequencing were preferentially added to the reference database rather than variant sequences from. COSMIC or dbSNP[23–25]; in some studies, after appending variants identified from sequencing data, the reference database was reduced to only proteins with transcript expressed according to transcriptomics, since according to the “central dogma of biology”, there can be no protein without corresponding transcript[26–28]. It was also proposed to generate reduced transcriptome-informed protein databases by barely reducing the reference database to proteins with expressed transcript, without including any novel sequence, with the only declared objective of increasing sensitivity in the identification of known sequences[26,28,29]. These works claim that searching such reduced transcriptome-informed databases allows increasing the number of valid identifications. A strategy was also proposed to optimize the balance between lost identifications due to the

incompleteness of an excessively reduced database and additional identifications observed from searching reduced transcriptome-informed databases, so as to maximize the number of valid identifications[30]. However: *i.* Only limited attention was given to the mechanistic explanation of the increased number of identified peptides with respect to database size; *ii.* Little is known about the impact of reduced transcriptome-informed database searches on protein inference, in terms of ambiguity of protein identification and shared peptide assignments. Therefore, in this work, we investigated the use of reduced transcriptome-informed sample-specific protein databases, focusing on these two methodological aspects. Our investigations result into three conclusions. First, the reported increment of the number of identifications obtained searching reduced transcriptome-informed databases is a statistical artifact: it is only the spurious consequence of an underestimated FDR, which results from the TDC sensitivity to the database size (also reported in [31]). In other words, reducing the search database to increase sensitivity broadly amounts to validate peptide identifications at an FDR that is larger than reported, hereby questioning the validity of peptides only identified thanks to a reduced database search and comparability between studies. Second, searching reduced transcriptome-informed protein databases followed by accurate FDR control remains nonetheless of interest, for it decreases ambiguity of protein identifications by reducing the proportion of shared peptides and the size of protein groups. Finally, searches against the reference database followed by post-hoc filtering of proteins with no evidence of transcript expression provides comparable proteomic identifications to searches against the reduced transcriptome-based database, while guaranteeing more transparency and comparability between studies. We therefore provided an R code to perform post-hoc filtering on proteomic identifications based on transcript evidence and to manually inspect protein identifications and shared

peptide assignments within protein groups of interest together with transcript expression information.

RESULTS

Reduced transcriptome-informed database search does not increase sensitivity if FDR is accurately controlled

To investigate the impact of reduced transcriptome-informed protein databases on proteomic identifications, we used two human samples (hereafter referred to as Jurkat and Lung) for which matched transcriptome and proteome were publicly available (**Supp. Table 1**). For each of them, we built a sample-specific reduced transcriptome-informed protein database, in the following way. First, we processed transcriptome datasets and identified the set of transcripts expressed in each sample using StringTie, a common transcriptome assembly method (see “Transcriptome analysis” in Methods). Then, we generated reduced databases for MS/MS search by retaining from the Ensembl human protein database only those proteins whose transcript was expressed in the sample-matched transcriptome (see “Construction of reduced transcriptome-informed protein databases for MS/MS search” in Methods) (**Figure 1A-B**). We compared valid peptide-spectrum matches (PSMs) obtained from the MS/MS search against the Ensembl human database (referred to as the “full database”) or against the sample-specific reduced database (referred to as the “reduced database”) at 1% FDR, as estimated by TDC. In agreement with previous studies[26,28,30], we found that a few spectra and peptides identified in the full database were lost in the reduced database search (“lost in reduced DB”), while others were only identified in the reduced database search (“additional in reduced DB”)

(**Figure 1C**). Lost identifications originate from reduced database incompleteness. Indeed, even more identifications were lost when using a further reduced protein database, as the one generated on the basis of the smaller set of expressed transcripts identified by Cufflinks, an alternative method of transcriptome assembly (**Supp. Fig. 1**). Additional identifications, instead, are commonly attributed to an increased sensitivity of MS/MS searches against smaller databases, such as reduced transcriptome-informed protein databases[26,28,30]. In this work, we investigated more thoroughly the origin of additional identifications obtained from searching these reduced transcriptome-informed protein databases. To this end, we performed a detailed comparison of all (target or decoy) PSMs retained from the full and reduced database searches, after validation prefilters (*i.e.*, single best scoring PSM per spectrum, minimum peptide length of 7 amino acids), but prior to filtering for 1% FDR control (**Figure 1A**). Since we built the reduced database as a simple subset of the Ensembl human database and considered only a single best scoring peptide per spectrum (see “Proteome analysis” in Methods), we could easily map each spectrum match between the two searches. Two interesting observations emerged from this comparison. First, several spectra are reallocated in the reduced database (*i.e.*, assigned in the reduced database search to a different match from that of the full database), which occurs when the peptide match from the full database is not included in the reduced database. However, the PSM score is never higher in the reduced database: at best, it is equal (data points on the diagonal, **Figure 2A, Supp. Fig. 2A**) to the score in the full database, but for the most part it is smaller (**Figure 2B**, **Supp. Fig. 2B**). Therefore, additional identifications in the reduced database do not come from an improved search score. For sake of clarity, those few spectra with a match only in the reduced database (indicated as “no match, target” or “no match, decoy” in **Figure 2A, Supp. Fig. 2A, 3A, 4 and 5**) do not contradict this observation; they are all explained by reallocation and

prefilters used for validation (**Supp. Fig. 3B,** see **Supplementary Note 1** for a detailed explanation). Similarly, we also verified all cases of pretty rank (PSMs with score difference > 0.1, which are considered of equal score, see “Proteome analysis” section in Methods) and again confirmed that the maximum PSM score in the reduced database in no case is higher than in the full database search (**Supp. Fig. 3C-D**, see also **Supplementary Note 1**). The second main observation was that the score cutoff estimated by TDC at 1% FDR (*i.e.*, the score defining the set of accepted PSMs, while respecting the constraint that less than 1% of them are expected to be a false discoveries) is lower for the reduced database than for the full database search (**Figure 2A, Supp. Fig. 2A**). Consequently, for a few spectra their match is not validated after FDR control in the full database while, at lower or equal score at best, it is validated in the reduced database search (pointed out by the arrow in **Figure 2A** and **Supp. Fig. 2A**). This is clearly the reason why these PSMs are accounted for as additional identifications in the reduced database search. We also observed a few reallocations, which can likewise yield additional spectra and/or peptide identifications in the reduced database search. They are, in particular, reallocations from non-target matches in the full database to target matches in the reduced database search (2.9% and 1.9% of all spectra in Jurkat and lung respectively) and reallocations between different target matches (3.2% and 2.7% of all spectra in Jurkat and lung respectively) (**Figure 2A,C and Supp. Fig. 2A,C**). However, only a minority of them are valid identifications at 1% FDR control (*i.e.*, pass the score cutoff for FDR control from the reduced database search) (**Figure 2C, Supp. Fig. 2C, Supp. Table 4**). Further, even fewer of them would pass the cutoff obtained from the full database search and they are hereafter named as “pure reallocations” to indicate that additional identifications from these PSMs uniquely originate from reallocation and not from additionally validated PSMs due to the the lower cutoff at 1% FDR validation in the reduced database (**Figure 2C**,

Supp. Fig. 2C and 6A, Supp. Table 4). Additional peptide identifications originating from either lower cutoff for FDR control or from pure reallocations present inferior PSM scores compared to peptide identifications obtained from both database searches (**Supp. Fig. 6B**). For pure reallocations, the difference in score between the full and reduced database match can be quite important, especially for target PSMs in the full database, reallocated in the reduced database search (**Supp. Fig. 6C**). Furthermore, additional peptide identifications only allow obtaining 6 and 8 additional protein identifications (*i.e.*, protein groups whose protein members are not identified in the full database search) in the Jurkat and lung sample, respectively (**Supp. Table 5**). Thus, these additional identifications are of lower quality and provide little benefit to protein identification. Overall, only few additional identifications come from pure reallocations, while the main origin is the lower cutoff for FDR control in the reduced database, explaining 98.6% (n=3,147) and 95.2% (n=1,875) additional spectral identifications and 96.5% (n=5,560) and 77.5% (n=1,524) additional peptide identifications for the Jurkat and lung samples, respectively (**Figure 2D-E, Supp. Fig. 2D-E**). Therefore, we investigated the reasons why lower cutoffs were observed at a same FDR threshold in the reduced databases. To do so, we first simulated what would occur if it were instead equal to that of the full database (**Figure 3A, Supp. Fig. 7A**). We observed that the proportion of valid decoys in the reduced database search would considerably decrease compared to the full database, with a net loss of 38.4% and 27.1% of valid decoys in the Jurkat (**Figure 3B**) and lung sample (**Supp. Fig. 7B**), respectively. Indeed, an important fraction of spectra matching valid decoys in the full database are assigned to invalid or non-decoy matches in the reduced database and not counterbalanced by reallocations in the other direction (*i.e.*, from invalid/non-decoy matches to valid decoys) (**Figure 3C, Supp. Fig. 7C, 4 and 5**). On the contrary, the majority of spectra matching valid targets in the full database match the same valid target

in the reduced database, so that their loss is quite limited (**Figure 3C, Supp. Fig. 7C**). While spectra matching valid decoys in the full database are reallocated much more frequently in the reduced database than spectra matching valid targets, upon reallocation they behave similarly: only few reallocations result in a valid match of the same type (**Figure 3C**, **Supp. Fig. 7C**) and the score difference between full and reduced database matches is comparable (**Supp. Fig. 8A**). Hence, the proportion of valid decoys lost in the reduced database is higher than that of targets, simply because a higher proportion of them is reallocated. This is easily explained by how the reduced database is generated: only proteins whose transcript is expressed, thus those more likely to be present, are retained from the canonical full protein database. Therefore, all valid targets from the full database are in theory still present in the reduced database, while this is not the case for decoys that represent by definition random matches (**Supp. Fig. 8B**). The lower cutoff obtained by TDC for the reduced database allows to validate a few more decoys and thus recover the proportion of valid decoys required to declare a nominal FDR level of 1% (**Figure 3B, Supp. Fig. 7B**). We claim that additional identifications validated using a lower cutoff in the reduced database represent a byproduct of the known influence of the database size on TDC[31], rather than an effect of increased sensitivity in reduced database searches. Naturally, in absence of a benchmark, it is impossible to determine whether they represent correct matches, missed in the full database due to FDR overestimation, or incorrect matches, accepted in the reduced database due to FDR underestimation. However, three main observations indicate that they should be at least considered with caution. First, they are accepted in the reduced database at quite low scores, meaning that, in any case, they represent low quality spectra and cannot be identified with very high confidence. Second, it could be assumed that additional identifications stem from the removal, in the reduced database, of high-scoring decoys out-competing correct target

matches, thus lowering sensitivity. However, in our study, most additional identifications do not represent reallocations from decoys to targets; they consist, instead, in the same PSMs, accepted at 1% FDR only in the reduced database because of a lower score cutoff (**Figure 2D-E; Supp. Fig. 2D-E**). Third and most importantly, the artifactual origin of additional identifications, given TDC sensitivity to database size, is supported by the comparison between the behavior of TDC and of the Benjamini-Hochberg (BH) procedure for FDR control[32]. BH is known to be a conservative and stable FDR control procedure, and it has recently been successfully applied to peptide identification[31]. In particular, TDC was found to be less conservative and less stable than BH with respect to preliminary filters on precursor mass accuracy: at narrower mass tolerance, fewer decoys were fair competitors for incorrect random matches, artificially lowering cutoffs. Therefore, reducing database size can similarly result into an insufficient number of decoys to accurately simulate incorrect target matches and lead to the observed lower cutoffs. To confirm this, we applied the BH procedure on target-only database searches (see “Proteome analysis” section in Methods) and obtained more conservative score cutoffs and, most importantly, more stable with respect to database size, compared to TDC (**Figure 3D, Supp. Fig. 7D, Supp. Fig. 9A-B**). Consistently, a much more limited number of additional identifications was validated in reduced database searches using BH-based FDR control (**Supp. Fig. 9C**). We also employed the BH procedure for FDR control on concatenated target-decoy database searches; while doing so is a nonsense from a practical data processing viewpoint, yet from a statistical methodology viewpoint, it simplifies comparative evaluations of BH and TDC stabilities. As expected, we obtained more conservative and stable score cutoffs with BH (**Supp. Fig. 9A-B)**.

Transcriptome information aids to reduce ambiguity of protein identifications

While not enhancing sensitivity of peptide identifications, reduced transcriptome-informed databases can still benefit proteomics at the protein inference step, by lowering ambiguity in protein identifications. These databases include fewer proteins – only those proteins which are most likely present given their transcript expression – and it is reasonable to assume that with fewer possible protein matches, we may obtain fewer shared peptides and smaller protein groups. This decrement in protein group size has already been observed, but either not discussed[33] or attributed to an additional number of identifiable peptides available for parsimony-based protein inference[26]. We already illustrated how additional identifications from searching reduced databases actually represent a flaw of TDC with respect to reduced database size, and how they can be largely avoided using alternative procedures for FDR control, such as, for instance, BH. We will now show that plain searches against reduced databases followed by BH-based FDR control nonetheless yield smaller protein groups and less ambiguous protein identifications, thus regardless of additional identifications or protein inference methods. Concretely, we compared identifications obtained from the full or reduced database searches followed by BH-based FDR control. The total number of identifications, at the spectrum, peptide and protein level, is comparable (**Figure 4A**). As number of protein-level identifications, we used the number of protein groups, as defined by the Proline software, which include both the unambiguous identification of a single protein (single-protein groups) and groups of indiscernible proteins identified by the same sets of peptides (multi-protein groups) (see “Proteome analysis” in Methods). Interestingly, the proportion of single-protein groups is considerably higher for the reduced database (**Figure 4B**), meaning that protein identifications are less ambiguous.

We further characterized ambiguity of protein identifications using the graph’s connected components. Briefly, we first represented peptide-to-protein mappings via bipartite graphs,

with peptides and proteins as vertices and with edges featuring peptide to protein membership: this allows an easy picturing of the complex structures generated by shared peptides, as well as their processing by means of graph theory. Then, we calculated connected components (CCs), *i.e.*, the largest subgraphs in which any two vertices are connected to each other by a path and not connected to any other of the vertices in the supergraph. Proteins sharing one or more peptides are thus gathered in the same CC (multi-protein CCs), while unambiguous protein identifications are represented by CCs with a single protein vertex (single-protein CCs) (**Figure 4C)**. As such, CCs constitute a peptide-centric strategy to represent ambiguous protein identifications and their shared peptides, not to be confused with the classical protein-centric strategy of protein grouping. It presents two main advantages. First, it provides a non-redundant representation of shared peptides, which can instead be duplicated between different protein groups. Second, it is independent from the different existing strategies of protein inference and protein grouping, making it widely applicable, reproducible and transparent. We observed that, while the total number of obtained CCs is comparable, there is a considerably higher proportion of single-protein CCs in the graph derived from the reduced database search results. After the reduction of the protein group size, this is the second evidence of a decreased ambiguity of protein identifications (**Figure 4D)**. Consistently, we also observed a greater proportion of specific peptides – and a correspondingly lower proportion of shared peptides-- from the reduced database search (**Figure 4D)**. Within multi-protein CCs, the ratio between the number of protein members and the corresponding number of their encoding genes is also inferior for the reduced database, suggesting that at least part of the solved ambiguity occurred between proteins encoded by different genes (**Figure 4E, Supp. Table 6**). As a side note, we additionally observed that searches against reduced databases are associated with inferior ambiguity at the PSM level, although to a less

extent. In the peptide identification step, it is common to only consider the best peptide match for each spectrum (*i.e.*, the rank 1 PSM, according to the search engine score) but it can occur that a spectrum matches different peptides equally well (or almost equally). This complicates the analysis and no consensus exists on how to treat these cases. Interestingly, we observed that a smaller proportion of spectra with multiple best matches occur in the reduced database search (**Supp. Fig. 10A**); likewise, fewer best matches are in general found per spectrum (**Supp. Fig. 10B**).

At last, we adopted an alternative strategy to enhance proteomics by transcriptomics, which consists in an MS/MS search against the full database, followed by post-hoc filtering of proteins with no expressed transcript and no specific peptide (**Figure 5A, Supp. Fig. 11**). The driving principle is indeed to remove ambiguous protein identifications not supported by specific peptides or by transcriptomics and thus reduce ambiguity due to shared peptides. Overall, we observed similar results from reduced transcriptome-informed database searches and post-hoc filtering. First, they provide a similar number of spectra and peptide identifications, comparable to that of the full database search (**Figure 5B**); secondly, they yield a similarly higher proportion of single-protein CCs and specific peptides than full database searches (**Figure 5C**), indicating less ambiguous protein identifications. Post-hoc filtering is a transparent and easily interpretable approach and we believe that it is most suitable to studies aiming to enhance protein inference. While a few software tools already exist to generate reduced protein databases, we provide here a specific toolbox of R scripts to perform the aforementioned post-hoc filtering. The toolbox additionally allows very efficient calculation of the CCs, which we have proposed as a means to quantify and compare ambiguity of protein identifications, to visualize CCs of interest and manually inspect them before and after post-hoc filtering.

DISCUSSION

In this work, we provide guidance for a mindful use of reduced transcriptome-informed protein databases for MS/MS search. This type of reduced databases stems from the attempt to counter excessive database inflation in proteogenomic studies, when adding variant or novel proteoforms identified from sequencing data. Indeed, increased database size complicates the task of discriminating between correct and incorrect matches. When using TDC-based FDR control, inflated target databases come with an inflated number of decoys and consequently higher probability to get high-scoring decoy matches. This has mainly been thought to reduce sensitivity of identifications in two ways. First, decoy matches may score better than correct target matches and outcompete them in spectrum-peptide assignment ("outcompeting decoys"), so that the number of obtained identifications decreases. Second, decoys may have higher probability to be matched than incorrect targets, which violates the Equal Chance assumption of TDC procedure and provides an overestimated FDR, again decreasing the number of identifications. As the main *raison d'être* of proteogenomics is to maximize the number of identifiable peptides, including variants or non-canonical ones, many efforts were made to avoid loss of sensitivity from excessively large databases, for example by reducing their size. While issues coming from use of excessively large databases have been abundantly discussed, fewer works pointed out that also excessively small databases may be problematic, as they also affect TDC estimations[31,34,35]. With excessively small databases, TDC provides inaccurate FDR estimates, since these estimates can only be asymptotically accurate[34,36,37]. Further, with too few (high-scoring) decoys, the probability to match a

decoy may be lower than the probability to match an incorrect target, which violates again the Equal Chance assumption of TDC, leading this time to FDR underestimation and to an artifactual increase of identifications. In this work, we showed that the increment of identifications obtained by searching reduced transcriptome-informed protein databases is likely to represent a statistical artifact from employing TDC on excessively small databases. We illustrated how TDC estimates a lower score cutoff for 1% FDR control on the reduced databases compared to the full database search results, causing some invalid PSMs in the full database to be retained as valid additional identifications only in the reduced database. We confirmed this observation at various levels of FDR control (0.5%, 1%, 5%) and in two different human-derived samples – a tissue (lung) and a cell line (Jurkat) – with a different level of proteomic complexity and number of spectra. Fewer spectra are available for the Jurkat sample, which, interestingly, also presents a more important difference in score cutoffs between full and reduced database. Indeed, not only reduced database sizes but also a lower number of spectra is believed to affect TDC ability to accurately estimate FDR[34]. We claim that additional identifications obtained from such reduced databases are at least doubtful and that it is unwise to employ reduced transcriptome-informed protein databases with the aim of increasing the number of identifications. Indeed, the obtained additional identifications have quite low scores and do not stem from removal of out-competing decoys, a known cause of missed identifications in excessively large databases; instead, they rather represent PSMs identical in the two database searches but only accepted in the reduced database due to a lower score cutoff for the same level of FDR control. Most importantly, only a negligible number of additional identifications is generated from the reduced database search when using a method for FDR control known to be stable with respect to database size, such as BH. Indeed, using BH, score cutoffs estimated for the full and reduced database searches, at the same level

of FDR control, are almost identical. As is, BH procedure constitutes an interesting alternative to TDC for stable FDR control irrespective of the database size. However, many alternative approaches have been recently developed to cope for the weaknesses of classical TDC[38–43]. It is important that proteogenomics researchers use them to avoid risking statistical artifacts in their data. In doing so, they will not benefit any longer from the so far hypothesized sensitivity increment, but this seems to be the necessary cost for rigorous control of the FDR.

Reduced transcriptome-informed protein databases are nonetheless useful in bottom-up proteomics to reduce ambiguity of protein identifications, which comes from the presence of shared peptides. In particular, we showed that searching these reduced databases yields a higher proportion of specific peptides and unambiguously identified proteins (*i.e.*, single-protein CCs). Furthermore, the higher proportion of specific peptides and correspondingly lower proportion of shared peptides positively affects precision in relative protein quantification. Indeed, in relative protein quantification, where peptide abundances are used as a proxy for the abundance of their parent protein, shared peptides are difficult to handle: since their relative abundance may depend on the contribution of multiple proteins they are frequently discarded. As a downside, this heavily restricts the number of remaining quantifiable proteins, which is reduced to proteins with at least one specific peptide, and the amount of information available to estimate abundances, corresponding to the number of specific peptides only. Therefore, a lower proportion of shared peptides represents more information available for quantification.

Finally, we showed that full database searches followed by post-hoc filtering of proteins with no expressed transcript provide proteomic identifications comparable to reduced database searches and similarly reduces ambiguity of protein identifications, while being more transparent and interpretable. We provided an R code to implement such

post-hoc filtering strategy. The code allows the user to visualize ambiguous protein identifications and their peptides via bipartite graphs, to prune them according to transcript expression and to manually inspect how this transcriptome-based post-hoc filtering strategy reduces ambiguity. Ambiguous protein identifications are represented and quantified using graph connected components, which constitute here subgraphs of proteins connected by shared peptides. This representation comes with the following advantages: it is transparent, interpretable, non-redundant with respect to shared peptides and independent from the variety of different strategies developed to define protein groups.

Results from this work are of interest also beyond proteogenomics. Indeed, database reduction is widely pleaded in proteomics, while little attention is being paid to the limitations of TDC when using excessively small databases. It was proposed for example, to limit database size based on peptide detectability[30] and it was more generally claimed that “mass spectrometrists should only search for peptides they care about”[44]. The observed TDC statistical artifacts with excessively small databases is an issue similarly concerning multi-step search strategies implemented by some proteomics search engines[45] or developed in metaproteomics[46,47]. Furthermore, the observation that transcriptome information can aid to decrease ambiguity in protein identification is of general relevance in classical proteomics and even more in metaproteomics, which is confronted with an additional source of protein ambiguity, namely the presence of multiple organisms in the same sample.

METHODS

Proteogenomic datasets description

We analysed two samples for which matched transcriptome and proteome were publicly available: a healthy lung tissue and a Jurkat cell line. The lung sample comes from a dataset by Wang *et al.*[33], which includes 29 histologically healthy human tissues and was meant to describe mRNA and protein expression levels across human body. The lung transcriptome dataset was obtained by paired-end RNA sequencing on an Illumina HiSeq 2000/2500 system generating 2×100 bases long reads. Its matched proteome dataset was obtained by quantitative label-free LC-MS/MS using an on-line nanoflow liquid chromatography system coupled to a Q Exactive Plus mass spectrometer, operating in data-dependent mode. Sample preparation included peptide fractionation via hSAX (hydrophylic strong anion) chromatography. Transcriptome and proteome raw data were downloaded from the EBI SRA (ArrayExpress accession: E-MTAB-2836; run accession: ERR315346) and the ProteomeExchange (dataset identifier: PXD010154; sample identifier: P013163) repositories, respectively.

The Jurkat cell line dataset comes from a study by Sheynkman *et al.*[24]. The Jurkat transcriptome dataset was obtained by paired-end RNA sequencing on an Illumina HiSeq 2000 system generating 2×200 bases long reads. The matched proteome dataset was obtained by quantitative label-free LC-MS/MS using nanoAquity LC system chromatography system coupled to a Velos-Orbitrap mass spectrometer, operating in data-dependent mode. Sample preparation included peptide fractionation via high pH LC separation. Transcriptome and proteome raw data were downloaded from the NCBI's

Gene Expression Omnibus (GEO) and the PeptideAtlas repositories with accession GSE45428 and PASS00215, respectively.

Transcriptome analysis

Raw reads were downloaded from public repositories and processed on the Galaxy platform available at https://usegalaxy.org/[48] using common workflows of read preprocessing and alignment for transcript identification (**Supp. Table 3**). First, sequencing adapters and low quality (Phred score < 20) read ends were trimmed off using TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and reads shorter than 20 bp after trimming were discarded. Then, preprocessed reads were aligned against the human reference genome (assembly GRCh38) by the splice-aware STAR aligner[49] in default mode, using the Ensembl reference gene model for splice junctions. Only reads mapped in a proper pair, passing platform quality checks were retained. Reads corresponding to optical or PCR duplicates were removed, as well as non-primary and supplementary alignments. We initially employed two common strategies of transcriptome assembly and quantification: StringTie[50] and Cufflinks[51]. Both programs were run looking for reference transcripts only (no novel transcripts were searched) and they yielded two comparable set of expressed transcripts. Unless otherwise specified, StringTie output was used for downstream analyses.

Construction of reduced transcriptome-informed protein databases for MS/MS search

For each sample, we built sample-specific protein databases for MS/MS search, containing only those protein sequences for which the corresponding transcript is

expressed in the sample. Briefly, we first processed sample-matched transcriptomes as described above and identified the subsets of transcripts expressed at FPKM>1 according to the StringTie or Cufflinks algorithms of transcript assembly and quantification. Then, we filtered the human GRCh38 Ensembl protein database, only keeping those proteins whose corresponding transcript is expressed in the sample. For each sample, we obtained two sample-specific reduced versions of the Ensembl database, based on expressed transcripts from either StringTie or Cufflinks transcript quantification (**Supp. Fig. 1A**). Unless otherwise specified, all downstream analyses were performed using the reduced transcriptome-informed database built according to expressed transcripts identified by StringTie, which is more recent than Cufflinks.

Proteome analysis

Raw spectra were downloaded from public repositories and processed automatically using Mascot Distiller software (version 2.7, Matrix Science). Peptide spectrum matches were identified using Mascot search (version 2.6) against two different concatenated target-decoy databases: either the original human GRCh38 Ensembl protein database (release 98, September 2019) or a reduced version of it containing only proteins whose transcript is expressed (as described in the “Construction of reduced transcriptome-based protein databases for MS/MS search”). In both cases an equivalent number of decoy sequences was appended, as well as a custom database of common contaminant sequences (n=500) (and the corresponding number of decoys). Decoy sequences were generated by reversing target sequences with the perl script provided with the Mascot software. The parameters used for Mascot search on the lung and Jurkat samples are reported in **Supp. Table 2**.

The Proline software[52] was used for post-search PSM validation with the following prefilters: *i.* PSMs with score difference < 0.1 were considered of equal score and assigned to the same rank (pretty rank); *ii.* only a single best-scoring PSM is retained per query (single PSM per rank); *iii.* minimum peptide length >= 7 amino acids. Prefiltered PSMs were then filtered at the score cutoff estimated for 1% FDR control. Unless otherwise specified, the score cutoff for FDR control was estimated by target-decoy competition[53]. No protein inference was performed but for each peptide, all possible protein matches were considered. Protein identifications were reported as protein groups, as defined in the Proline software. Protein groups include both the unambiguous identification of a single protein (single-protein groups) and groups of undiscernible proteins identified by the same sets of peptides (multi-protein groups).

Further analyses were performed using the Benjamini-Hochberg procedure for FDR control[32], in alternative to TDC (see results section “Transcriptome information aids to reduce ambiguity of protein identifiifcations”). For these analyses, we used PSMs obtained from target-only protein databases, appended with the same database of common contaminant sequences, and searched with the same Mascot parameters as before.

Peptide-protein bipartite graphs and connected components

We represented proteomic identifications using bipartite graphs with two types of nodes -- *i.* identified peptides; *ii.* all their possible proteins of origin -- to more easily analyze and visualize groups of ambiguous protein identifications connected by shared peptides. Indeed, peptide assignments to proteins may be very complex in presence of shared peptides, but are easily represented using bipartite graphs. We then employed graph connected components (CCs), defined as the largest subgraphs in which any two

vertices are connected to each other by a path and not connected to any other of the vertices in the supergraph, to quantify and visualize ambiguity of protein identifications.

To build bipartite graphs of proteomic identifications, we first generated a tab-separated file containing for each identified peptide all proteins it matches to (one per line), based on the output of PSM validation by the Proline software. We then converted it into an incidence matrix, with proteins along the columns and peptides along the rows, using the crosstab function from the GNU datamash program (http://www.gnu.org/software/datamash). By cross-product of the incidence matrix, we obtained the corresponding adjacency matrix, which describes protein-to-protein connections, based on shared peptides. Finally, we calculated CCs, using the connComp() function of the "graph" R package on the adjacency matrix. There are two types of CCs: *i.* those containing one single protein (single-protein CCs), with only specific peptides, which constitute unambiguous protein identifications; *ii.* those containing multiple proteins sharing peptides (multi-protein CCs), which represent ambiguous protein identifications. Ambiguous protein identifications can be visually inspected by taking the CC of interest, extracting from the incidence matrix all specific and shared peptides mapping on the CC protein members and plotting peptide-to-protein mappings as bipartite graphs, using the "*igraph"* R package.

To decrease the computational cost in case of very large datasets or scarce computational resources, we also developed an alternative strategy of CCs calculation (**Supp. Fig. 12**). First the incidence matrix is reduced by removing all proteins not sharing peptides and all peptides unique to these proteins. Then the corresponding adjacency matrix is generated by cross-product of the incidence matrix and connected components can be more rapidly calculated on this reduced adjacency matrix. In this case, we exclusively obtain multi-protein CCs, since protein identifications with only specific

peptides, which correspond to single-protein CCs, were first removed from the incidence matrix. While multi-protein CCs are those of interest when investigating ambiguous protein identifications from shared peptides, single-protein CCs can still be easily retrieved from the original incidence matrix if required.

A companion R code is provided. It implements all the above described steps, including: *i.* generating the adjacency matrix; *iii.* calculating connected components; *iii.* visualizing CCs as bipartite graphs.

Transcriptome-informed post-hoc filtering

As an alternative to searching a reduced transcriptome-informed database, we tested a transcriptome-informed post-hoc filtering strategy, which works as follows. First, peptide identifications are obtained searching the full reference protein database and validated using the Proline software, as already described in “Proteome analysis” section in Methods. An incidence matrix is generated to encode peptide-to-protein mappings (see “Peptide-protein bipartite graphs and connected components” section). Then, the sample-matched transcriptome is analysed to identify the set of expressed transcripts. Finally, the peptide-to-protein incidence matrix is filtered by removal of proteins with no expressed transcript and no specific peptide. This allows to reduce ambiguity of protein identifications without loosing any peptide identification. The one-to-one transcript-to-protein correspondence is guaranteed by the adoption of Ensembl as reference protein database in proteomics and as genome annotation in transcriptomics. The filtered incidence matrix is then converted to an adjacency matrix to calculate CCs, as previously described (see “Peptide-protein bipartite graphs and connected components” section).

In this work, we employed the post-hoc filtering strategy on PSMs obtained from searching the target-only full Ensembl protein database followed by Benjamini-Hochberg procedure for FDR control, for comparability with the approach searching target-only reduced transcriptome-informed protein databases, followed by Benjamini-Hochberg procedure for FDR control. Indeed, Benjamini-Hochberg procedure was used in alternative to TDC after searching reduced transcriptome-informed protein databases to obtain accurate FDR control, as explained in the result section (see section “Reduced transcriptome-informed database search does not increase sensitivity if FDR is accurately controlled”). However, in other contexts, post-hoc filtering can be equally well adopted using PSMs from concatenated target-decoy searches followed by TDC-based FDR control. An R code implementing this post-hoc filtering strategy is also provided. In addition to filtering, the code allows visualizing the CCs of interest both before and after post-hoc filtering.

ABBREVIATIONS

BH Benjamini-Hochberg procedure

CC connected component

DB database

FDR false discovery rate

MS/MS tandem mass spectrometry

PSM peptide-sepctrum match

TDC target-decoy competition

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All datasets analyzed in this study are publicly available data from the works of Wang et al.[33] and Sheynkman et al.[24].

The transcriptome datasets were downloaded from the EBI SRA with ArrayExpress accession: E-MTAB-2836 (run accession: ERR315346) and from the NCBI's Gene Expression Omnibus (GEO) with accession GSE45428. The proteome datasets were downloaded from PeptideAtlas with accession PASS00215 and from the ProteomeExchange repository with dataset identifier: PXD010154 (sample identifier: P013163).

The code used to perform analyses is available as an R package, under GPL-3 license, at https://github.com/laurafancello/CCs4prot. Guidelines for installation and usage are documented in the GitHub repository.

Acknowledgements

The authors would like to thank IFB (Institut Français Bioinformatique) infrastructure for providing computational resources and members of the Edyp team for useful comments and feedback.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

1. Jaffe, J. D., Berg, H. C. & Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4,** 59–77 (2004).
2. Ruggles, K. V *et al.* Methods , Tools and Current Perspectives in Proteogenomics. *Mol Cell Proteomics* **16,**959–981 (2017).
3. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–25 (2014).
4. Zhang, B., Chambers, M. C. & Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **6**, 3549–3557 (2007).
5. Yang, X. *et al.* DBParser : Web-Based Software for Shotgun Proteomic Data Analyses. *J. Proteome Res.* **3**, 1002–1008 (2004).
6. Li, Y. F. *et al.* A bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* **16**, 1183–1193 (2009).
7. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
8. Serang, O., MacCoss, M. J. & Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **9**, 5346–5357 (2010).

9. Huang, T. & He, Z. A linear programming model for protein inference problem in shotgun proteomics. *Bioinformatics* **28**, 2956–2962 (2012).

10. Shanmugam, A. K., Yocum, A. K. & Nesvizhskii, A. I. Utility of RNA-seq and GPMDB Protein Observation Frequency for Improving the Sensitivity of Protein Identification by Tandem MS. *J. Proteome Res.* **13**, 4113-9 (2014).

11. Ramakrishnan, S. R. *et al.* Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics* **25**, 1397–1403 (2009).

12. Carlyle, B. C. *et al.* Isoform-Level Interpretation of High-Throughput Proteomics Data Enabled by Deep Integration with RNA-seq. *J. Proteome Res.* **17**, 3431–3444 (2018).

13. Bunger, M. K. *et al.* Detection and Validation of Non-synonymous Coding SNPs from Orthogonal Analysis of Shotgun Proteomics Data. *J. Proteome Res.* **6**, 2331–2340 (2007).

14. Alfaro, J. A. *et al.* Detecting protein variants by mass spectrometry: A comprehensive study in cancer cell-lines. *Genome Med.* **9**, 1–12 (2017).

15. Li, J. *et al.* A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. *Mol Cell Proteomics* **10,** 1–11 (2011).

16. Guerrero-Sanchez, V. M., Maldonado-Alconada, A. M., Sànchez-Lucas, R. & Rey, M.-D. Specific protein database creation from transcriptomics data in nonmodel species: Holm Oak (Quercus ilex. L.). *Methods Mol. Biol.* **2139**, 57–68 (2020).

17. Maringer, K. *et al.* Proteomics informed by transcriptomics for characterising active transposable elements and genome annotation in Aedes aegypti. *BMC Genomics* **18**, 101 (2017).

18. Blakeley, P. & Overton, I. M. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *J. Proteome Res.* **11**, 5221–5234 (2012).

19. Li, H. *et al.* Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **17**, (2016).

20. Krug, K., Popic, S., Carpy, A., Taumer, C. & Macek, B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *Proteomics* **14**, 2699–2708 (2014).

21. Park, H. *et al.* Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. *Proteomics* **14**, 2742–2749 (2014).

22. Verbruggen, S. *et al.* Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics Authors Spectral Prediction Features as a

Solution for the Search Space Size Problem in Proteogenomics. *Mol Cell Proteomics* **20**, 100076 (2021).

23. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).

24. Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-seq. *Mol. Cell. Proteomics* **12**, 2341–2353 (2013).

25. Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Scalf, M. & Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from non-synonymous nucleotide differences. *J. Proteome Res.* **13**, 228–240 (2014).

26. Wang, X. *et al.* Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **11**, 1009–1017 (2013).

27. Wang, X. & Zhang, B. customProDB : an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235–3237 (2013).

28. Sheynkman, G. M. *et al.* Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **15**, 1–9 (2014).

29. Proffitt, J. M. *et al.* Proteomics in non-human primates : utilizing RNA-Seq data to improve protein identification by mass spectrometry in vervet monkeys. *BMC Genomics* **18**, 1–10 (2017).

30. Shanmugam, A. K., Nesvizhski, A. I., Arbor, A. & Arbor, A. Effective leveraging of targeted search spaces for improving peptide identification in MS/MS based proteomics. *J. Proteome Res.* **14**, 5169–5178 (2015).

31. Coute, Y., Bruley, C. & Burger, T. Beyond Target − Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics. *Anal Chem* **92**, 14898-14906 (2020).

32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

33. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol*. **15**, e8503 (2019).

34. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. Target-decoy approach and false dscovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22**, 1111–20 (2011).

35. Keich, U., Kertesz-farkas, A. & Sta, W. Improved False Discovery Rate Estimation Procedure for Shotgun Proteomics. *J Proteome Res*. **14,** 3148-61(2015).

36. Levitsky, L. I., Ivanov, M. V, Lobas, A. A. & Gorshkov, M. V. Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach. *J. Proteome Res.* **397**, 6–10 (2017).

37. Burger, T. Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *J Proteome Res.* **17**, 12-22 (2018).

38. Keich, U., Tamura, K. & Noble, W. S. Averaging strategy to reduce variability in target-decoy estimates of false discovery rate. *J. Proteome Res.* **18**, 585–593 (2019).

39. Emery, K., Hasam, S., Noble, W. S. & Keich, U. Multiple Competition-Based FDR Control and Its Application to Peptide Detection. Preprint at http://*arxiv.org/abs/1907.01458v2* (2019).

40. Yi, X., Gong, F. & Fu, Y. Transfer posterior error probability estimation for peptide identification. *BMC Bioinformatics* **21**, 173 (2020).

41. Lin, A., Plubell, D. L., Keich, U. & Noble, W. S. Accurately Assigning Peptides to Spectra When Only a Subset of Peptides Are Relevant. *J. Proteome Res.* **20**, 4153–4164 (2021).

42. Ge, X., Chen, Y. E., Song, D., Mcdermott, M. & Woyshner, K. Clipper : p-value-free FDR control on high-throughput data from two conditions. Preprint at http://*biorxiv.org/content/10.1101/2020.11.19.390773v7.full* (2021).

43. Etourneau, L., Varoquaux, N. & Burger, T. Unveiling the links between peptide identification and differential analysis FDR controls by means of a practical introduction to knockoff filters. Preprint at http://*biorxiv.org/content/10.1101/2021.08.20.454134v1.full* (2021).

44. Noble, W. S. Mass spectrometrists should only search for peptides they care about. *Nat. Methods* **12**, 605–608 (2016).

45. Everett, L. J., Bierl, C. & Master, S. R. Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies. *J Proteome Res*. **9**, 700–707 (2010).

46. Jagtap, P. *et al.* A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**, 1352–7 (2013).

47. Cheng, K. *et al.* MetaLab : an automated pipeline for metaproteomic data analysis. *Microbiome*. **5**, 157(2017).

48. Afgan, E. *et al.* The Galaxy platform for accessible , reproducible and collaborative biomedical analyses : 2018 update. *Nucleic Acids Res*. **46**, W537–W544 (2018).

49. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

50. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–5 (2015).

51. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

52. Bouyssie, D. *et al.* Proline : an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics*. **36**, 3148–3155 (2020).

53. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. **4**, 207–214 (2007).

FIGURE LEGENDS

Figure 1. Comparison of proteomic identifications from the full reference database or the transcriptome-informed reduced database searches. A. Graphical representation of the two MS/MS search strategies we compared. MS/MS searches were performed using Mascot against either the reference human Ensembl protein database (full protein DB) or a subset of the reference database, generated based on transcript expression (reduced protein DB). PSMs were first validated using Proline software with the following prefilters: *i.* PSMs with score difference < 0.1 were considered of equal score and assigned to the same rank (pretty rank); *ii.* only a single best-scoring PSM is retained per query (single PSM per rank); *iii.* minimum peptide length >= 7 amino acids ; then they were filtered at the score cutoff estimated by target-decoy competition for 1% FDR control. **B.** Size and overlap of the reference human Ensembl protein database (full protein DB) and the sample-specific reduced transcriptome-informed protein databases (reduced protein DB). **C.** Number of spectra (on the left) or peptides (on the right) exclusively identified in the reduced database (“additional in reduced DB” in blue) or exclusively identified in the full database (“lost in reduced DB” in red) searches. The net difference between additional and lost identifications in the reduced database is also reported on top of each bar (“net”).

Figure 2. Additional identifications from the reduced database search mainly originate from a lower cutoff for 1% FDR control (Jurkat sample). A. Scatter plot comparing PSMs obtained from the full or reduced database searches. Each data point represents a spectrum: its corresponding PSM score in the full and reduced database searches is reported on the x and y coordinates, respectively. A color code is used to

represent the type of match (“target”, “decoy”, or “no match”) for each spectrum in the two searches. Score cutoffs obtained by TDC at 1% FDR are also shown as red and blue lines for the full and reduced database searches, respectively. The up-right insert zooms in on PSMs accepted at 1% FDR only in the reduced database, due to lower score cutoff at 1% FDR (black arrow pointing on the dash circle). **B.** Number of reallocated spectra whose score in the reduced database search is equal to that in the full database or lower. The score from searching the reduced database is never observed to be higher than the score from the full database. **C.** Stripchart reporting the PSM score in the reduced database for spectra undergoing reallocations. Only reallocations to target matches in the reduced database are shown. Reallocations are grouped based on the type of match for the same spectrum in the full and reduced database searches (<*match full DB*>_<*match reduced DB*>). The number and percentage of all spectra in each group is reported on the left. The number of reallocations passing the reduced database cutoff for 1% FDR control is shown in blue: they represent valid reallocations in the reduced database (“nb valid reallocations”). The number of reallocations which would pass the full database cutoff for 1% FDR control is shown in red: they represent additional valid identifications exclusively generated by reallocation, independent from the lower score cutoff, and are thus referred to as pure reallocations (“nb valid pure reallocations”). **D.** Bar plot representing the number of spectra (on the left) or the number of spectra identifying additional peptides (on the right) exclusively identified in the reduced database search due to: *i.* lower score cutoff at 1% FDR control in the reduced database search compared to the full database; *ii.* pure reallocation. The former are additional identifications from PSMs only passing the cutoff from the reduced database search and which would not be accepted based on the full database cutoff. It includes cases of identical PSMs in both searches (*no reallocation*, in black) and cases of reallocation from decoy (orange), target (gray) or no match (magenta)

in the full database search to target matches in the reduced database. Additional identifications from pure reallocation, instead, are those exclusively originated by reallocation, which would also pass the full database cutoff (*i.e.*, independent from the lower score cutoff effect). The Venn diagram on top of the additional peptides graph illustrates the corresponding non-redundant number of additional peptides (*i.e.*, peptides not identified in the full database search) identified from these spectra.

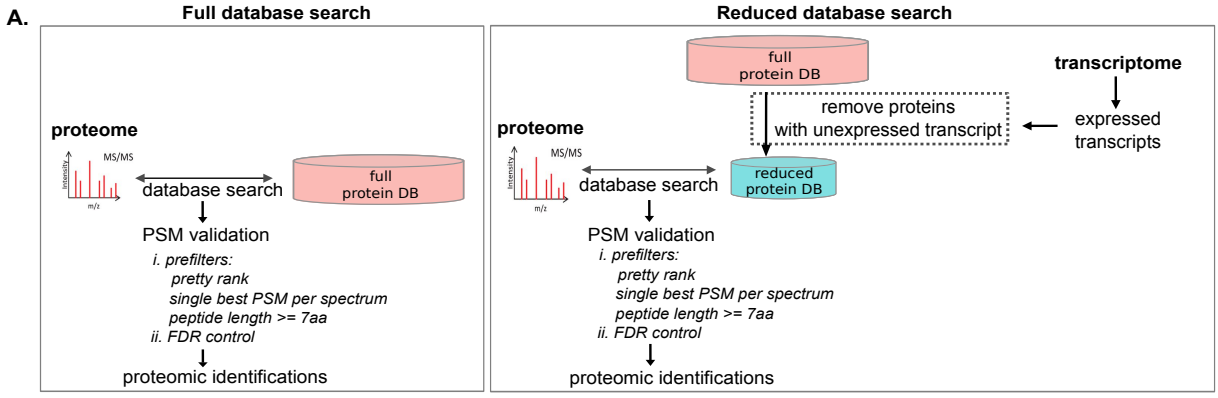
Figure 3. Lower cutoff for FDR control in the reduced database to recover valid decoys (Jurkat sample). A. Comparison of valid identifications obtained at 1% FDR from the full database (horizontal red arrow) or reduced database search (vertical blue arrow) and simulation of the valid identifications which would be obtained from the reduced database search if the score cutoff at 1% FDR were equal to that for the full database (dashed red arrow). **B.** Number of valid targets and decoys from the full or reduced database obtained at 1% FDR using the cutoffs estimated by TDC on the respective database search results (first and last rows). The second row instead simulates the number of valid targets and decoys which would be obtained from the reduced database if the estimated cutoff were the same as for the full database. Variations expressed in percentages are shown in gray. The associated nominal FDR level is reported (calculated as (d+1)/t, with *d* and *t* being the number of valid decoys and targets). **C.** Match in the reduced database search for spectra matching valid targets or valid decoys in the full database. **D.** Score cutoffs obtained by TDC or by BH procedure for FDR control for the full or reduced database searches at various FDR levels (0.5%, 1% and 5%). The variation of score cutoff between full and reduced database searches is reported in percentage.

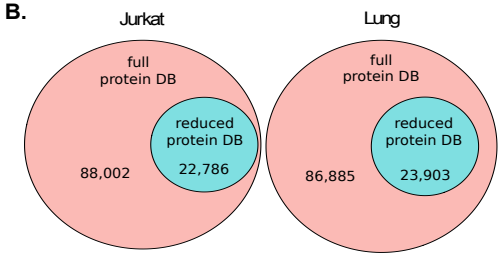
Figure 4. Transcriptome-informed reduced databases yield less ambiguous protein identifications. A. Number of valid identifications obtained from the full (red) or reduced (blue) target-only database searches, followed by BH procedure for 1% FDR control. The number of valid spectra, peptide and protein identifications are reported. Protein groups, as defined by the Proline software, represent here protein identifications and they include: *i.* proteins unambiguously identified by only specific peptides (single-protein protein groups); *ii*. groups of proteins identified by the same set of shared peptides (multi-protein protein groups). **B.** Percentage of single-protein groups. **C.** Bipartite graph representation of peptide-to-protein mappings and usage of graph connected components to visualize and quantify ambiguity of protein identifications. Unambiguous protein identifications are represented by CCs with a single protein vertex (single-protein CCs), while proteins sharing peptides are gathered in the same CC (multi-protein CCs) **D.** Upper panel: total number of connected components. Lower panel: percentage of specific peptides and of single-protein CCs. **E.** Genes encoding proteins from the full and reduced database searches. Upper panel: total number of genes associated to protein matches from the two searches. Lower panel: ratio between the number of protein members in each multi-protein CC and the number of their encoding genes.

Figure 5. Transcriptome-informed post-hoc filtering and reduced database search strategies similarly reduce protein identification ambiguity. A. Illustration of the transcriptome-informed post-hoc filtering strategy. First, an MS/MS search was performed against the full canonical protein database. Then, proteins with no corresponding expressed transcript in the sample-matched transcriptome and with no specific peptide (both conditions required) are removed, as well as peptides only mapping to that set of proteins. **B.** Number of valid spectra and peptide identifications obtained from the full or

884 reduced target-only database search (red and blue) or from the post-hoc filtering strategy
885 (orange), after 1% FDR control by BH procedure. **C.** Quantification of protein ambiguity for
886 the full or reduced database search (red and blue) or the post-hoc filtering strategies
887 (orange). Upper panel: total number of obtained CCs. Lower panel: percentage of specific
888 peptides and single-protein CCs.

Fig. 1


A.
Full database search
proteome
MS/MS
database search
full protein DB
PSM validation
i. prefilters:
pretty rank
single best PSM per spectrum
peptide length >= 7aa
ii. FDR control
proteomic identifications
Reduced database search
full protein DB
transcriptome
remove proteins with unexpressed transcript
expressed transcripts
proteome
MS/MS
database search
reduced protein DB
PSM validation
i. prefilters:
pretty rank
single best PSM per spectrum
peptide length >= 7aa
ii. FDR control
proteomic identifications


B.
Jurkat
full protein DB
reduced protein DB
88,002
22,786
Lung
full protein DB
reduced protein DB
86,885
23,903

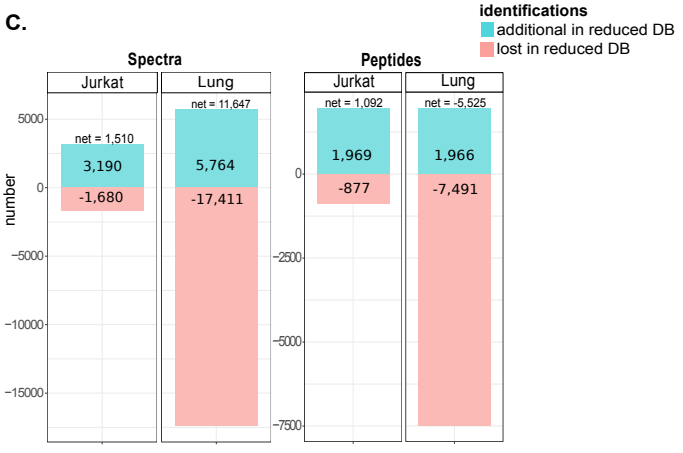
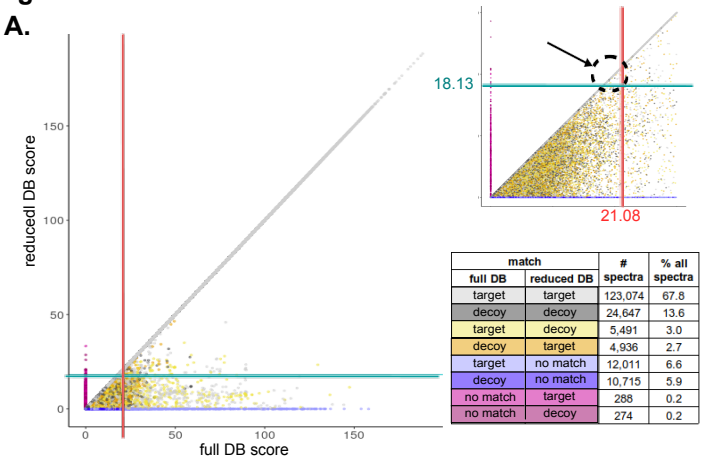
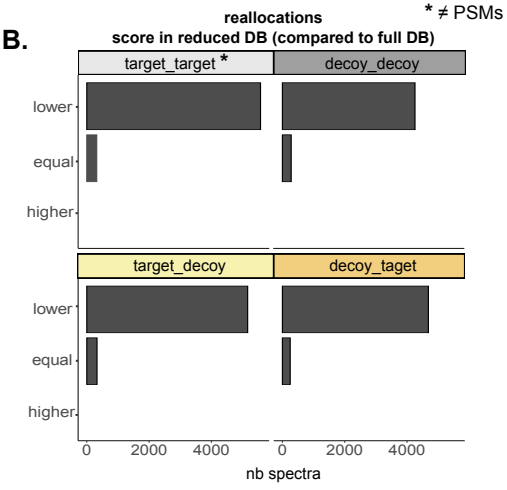
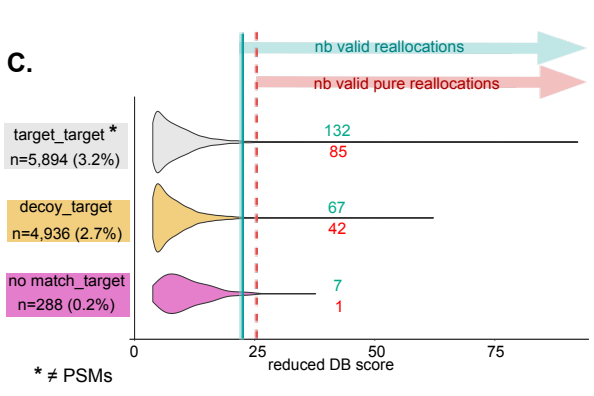

C.
identifications
additional in reduced DB
lost in reduced DB
Spectra
Peptides
Jurkat
Lung
number
net = 1,510
3,190
-1,680
net = 11,647
5,764
-17,411
net = 1,092
1,969
-877
net = -5,525
1,966
-7,491

Fig. 2



match		#	% all
full DB	reduced DB	spectra	spectra
target	target	123,074	67.8
decoy	decoy	24,647	13.6
target	decoy	5,491	3.0
decoy	target	4,936	2.7
target	no match	12,011	6.6
decoy	no match	10,715	5.9
no match	target	288	0.2
no match	decoy	274	0.2





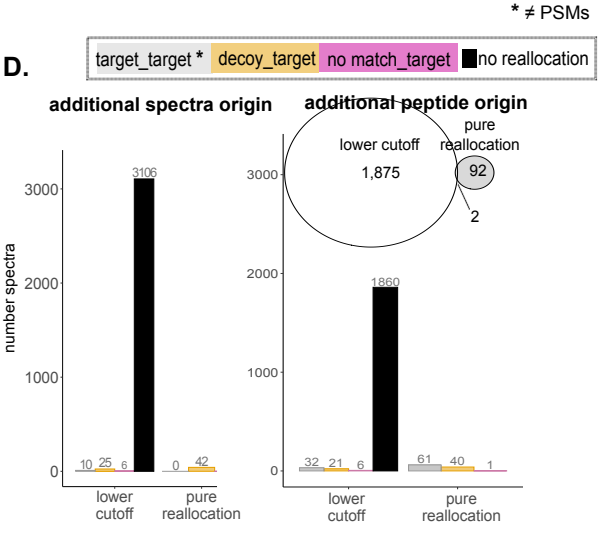
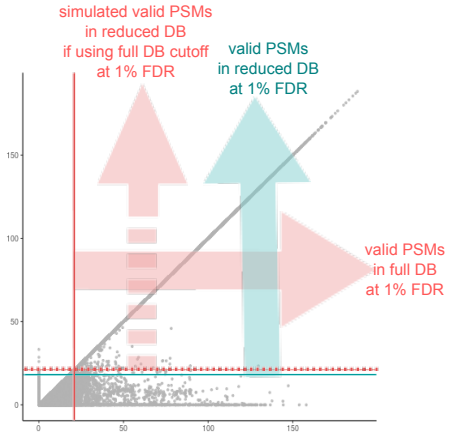


Fig. 3

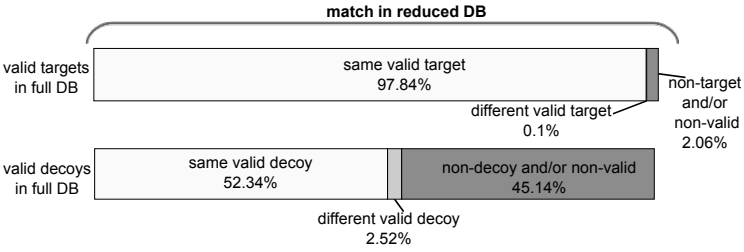
A.



B.

Database	Score cutoff	# valid targets	# valid decoys	FDR
full	21.08	83,455	833	1
reduced	21.08	81,781 -2%	513 -38.4%	0.63
reduced	18.13	84,965 +3.8%	847 +40.1%	1

C.



D.

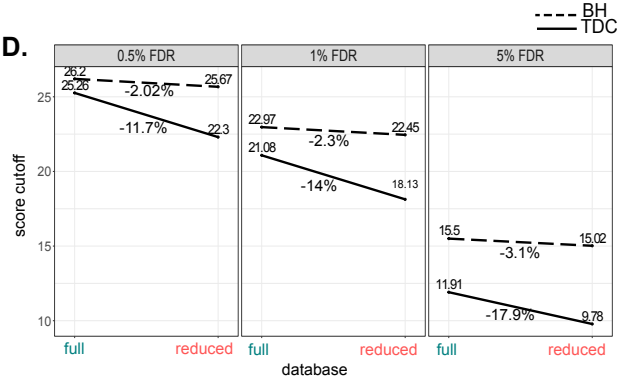
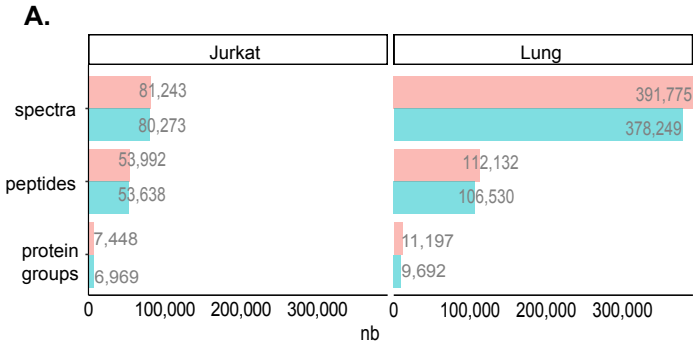
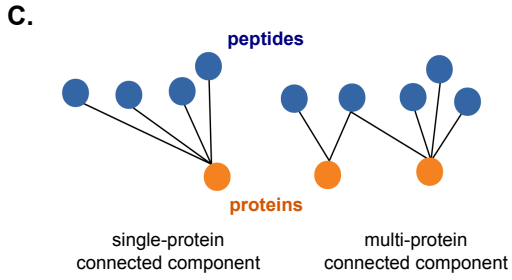
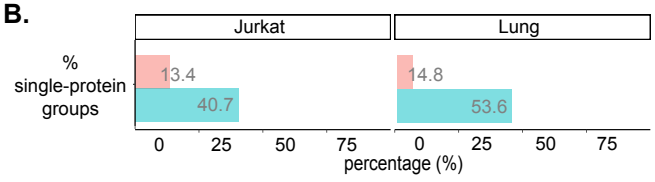
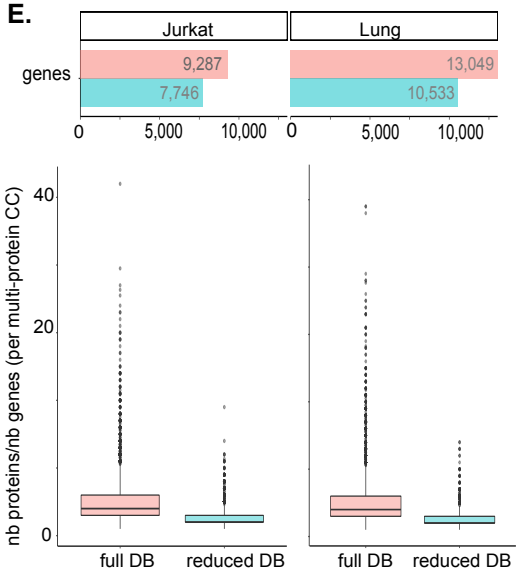


Fig. 4


A.
Jurkat
Lung
spectra
peptides
protein groups
81,243
80,273
391,775
378,249
53,992
53,638
112,132
106,530
7,448
6,969
11,197
9,692
0
100,000
200,000
300,000
nb


C.
peptides
proteins
single-protein connected component
multi-protein connected component


B.
Jurkat
Lung
% single-protein groups
13.4
40.7
14.8
53.6
0
25
50
75
percentage (%)


E.
Jurkat
Lung
genes
9,287
7,746
13,049
10,533
0
5,000
10,000
nb proteins/nb genes (per multi-protein CC)
40
20
0
full DB
reduced DB

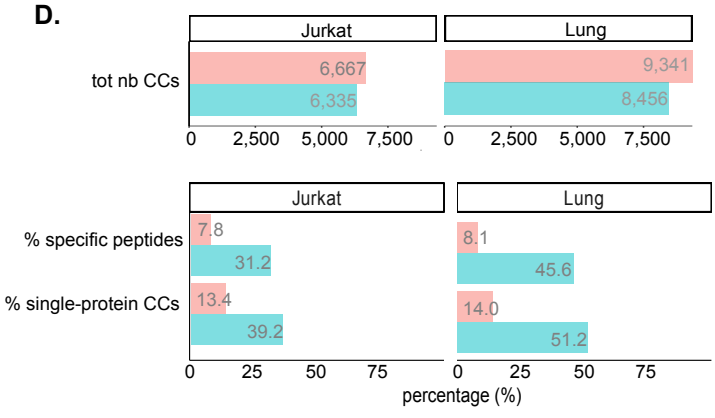
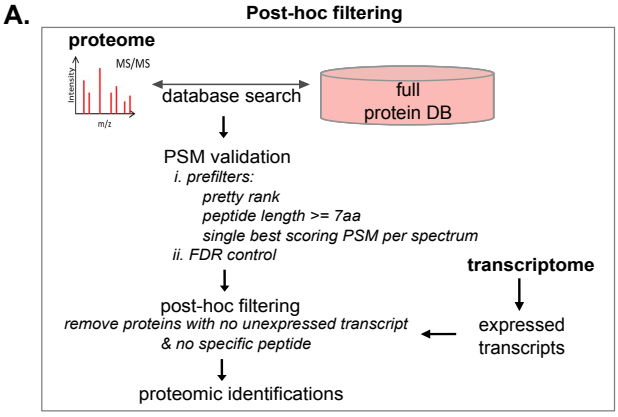
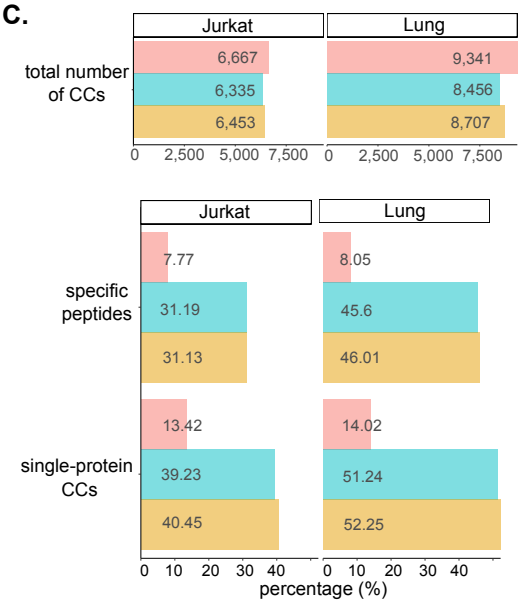
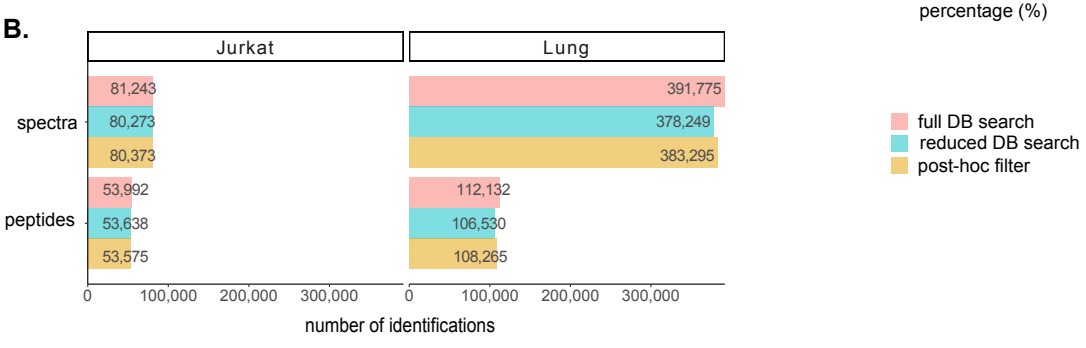

D.
Jurkat
Lung
tot nb CCs
6,667
6,335
9,341
8,456
0
2,500
5,000
7,500
% specific peptides
% single-protein CCs
7.8
31.2
8.1
45.6
13.4
39.2
14.0
51.2
0
25
50
75
percentage (%)

Fig. 5


A.
Post-hoc filtering
proteome
MS/MS
Intensity
m/z
database search
full
protein DB
PSM validation
i. prefilters:
pretty rank
peptide length >= 7aa
single best scoring PSM per spectrum
ii. FDR control
transcriptome
post-hoc filtering
remove proteins with no unexpressed transcript
& no specific peptide
expressed
transcripts
proteomic identifications


C.
Jurkat
Lung
total number
of CCs
6,667
6,335
6,453
9,341
8,456
8,707
0
2,500
5,000
7,500
specific
peptides
single-protein
CCs
7.77
31.19
31.13
8.05
45.6
46.01
13.42
39.23
40.45
14.02
51.24
52.25
0
10
20
30
40
percentage (%)


B.
Jurkat
Lung
spectra
peptides
81,243
80,273
80,373
391,775
378,249
383,295
53,992
53,638
53,575
112,132
106,530
108,265
0
100,000
200,000
300,000
number of identifications
full DB search
reduced DB search
post-hoc filter