Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics

Authors

Steven Verbruggen, Siegfried Gessulat, Ralf Gabriels, Anna Matsaroki, Hendrik Van de Voorde, Bernhard Kuster, Sven Degroeve, Lennart Martens, Wim Van Criekinge, Mathias Wilhelm, and Gerben Menschaert

Correspondence

Gerben.Menschaert@UGent.be

Graphical Abstract

In Brief

Proteogenomics suffers from statistical issues as the sequencing information inflates the database size. To compensate for this, rescoring with the machine learning–based spectrum predictors MS^{2}PIP and Prosit was implemented in a proteogenomics approach. This was demonstrated for both ribosome profiling and nanopore RNA-Seq derived databases. Postprocessing with Percolator showed that these techniques result in recovered and often even elevated stringency levels and identification rates. In this way, it allows to validate novel proteoforms through proteogenomics with unsurpassed confidence levels.
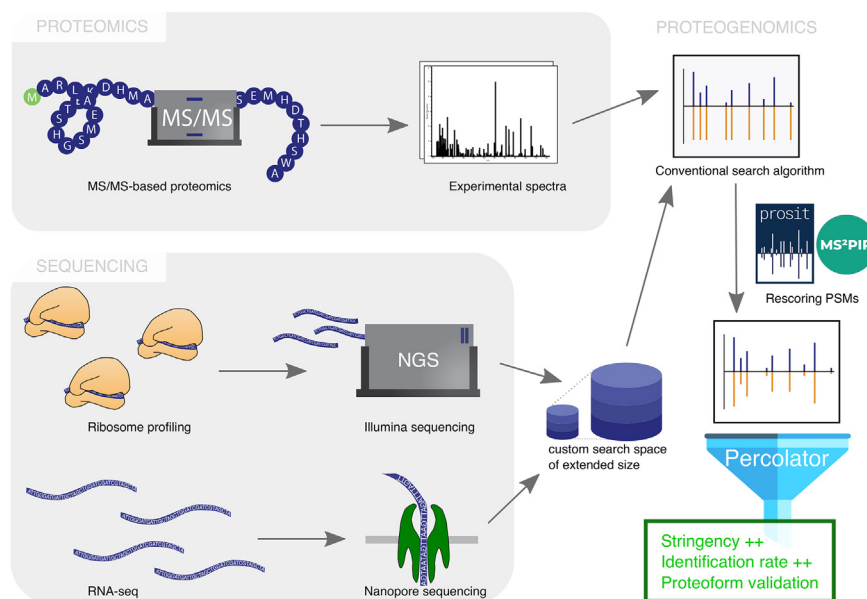
Highlights

- First proteogenomics with PSM rescoring using machine learning–predicted spectra
- Demonstrated on both ribosome profiling and nanopore RNA-Seq–derived databases
- Rescoring leads to elevated stringency and increased identification rates
- Rescoring compensates for the search space size issues in proteogenomics

Spectral Prediction Features as a Solution for the Search Space Size Problem in Proteogenomics

Steven Verbruggen[1,2], Siegfried Gessulat[3], Ralf Gabriels[4,5], Anna Matsaroki[2], Hendrik Van de Voorde[2], Bernhard Kuster[3], Sven Degroeve[4,5], Lennart Martens[4,5], Wim Van Criekinge[1], Mathias Wilhelm[3], and Gerben Menschaert[1,2,*]

Proteogenomics approaches often struggle with the distinction between true and false peptide-to-spectrum matches as the database size enlarges. However, features extracted from tandem mass spectrometry intensity predictors can enhance the peptide identification rate and can provide extra confidence for peptide-to-spectrum matching in a proteogenomics context. To that end, features from the spectral intensity pattern predictors MS²PIP and Prosit were combined with the canonical scores from MaxQuant in the Percolator postprocessing tool for protein sequence databases constructed out of ribosome profiling and nanopore RNA-Seq analyses. The presented results provide evidence that this approach enhances both the identification rate as well as the validation stringency in a proteogenomic setting.

Proteogenomics is defined as the research field in which proteomics is combined with genomics and/or transcriptomics. In practice, this is achieved by the generation of a custom protein sequence database from genomic or transcriptomic sequencing information, which can be subsequently used to identify nonreference peptides in mass spectrometry (MS) proteomics data (1). In this way, the protein-level validation, offered by proteomics, can be integrated with the depth and *de novo* capacities of sequencing technologies.

Nevertheless, proteogenomic approaches always imply an important trade-off: with increasingly more candidates in the custom search space, it gets more difficult to statistically differentiate true from false peptide-to-spectrum matches (PSMs). Two possible explanations for this effect are proposed. One describes that, with a growing number of candidates in the custom search space, it gets more likely that the best scoring PSM is actually a false one that jumped over a

true one by chance (2, 3). Another explanation states that a spectrum without a true match in the database has more options to produce a false PSM when the database is expanded. In this last case, the score distribution of the true PSMs does not change, but the score distribution of the false PSMs increases, which can result in true PSMs, priorly located at the false discovery rate (FDR) margins, now dropping below the raised FDR cutoff. Altogether, searching with larger databases can result in a number of novel peptide identifications, owing to the additional information available in the expanded search space. However, if the size of this database grows too drastically, the total number of identified peptides can drop considerably compared with proteomics performed with a conventional reference database (*e.g.*, UniProt), owing to the statistical issues explained above (3–5). In that case, the confidence gets underestimated owing to the database size (6).

A search space, constructed out of the complete *in silico* translated human genome, led to a 70-fold increase in size compared with the corresponding reference proteome from Ensembl (7). In order to narrow down this database size and at the same time use a more tailored approach, one can apply RNA-Seq on the same samples with the intention of deriving a sample-specific protein search space from the transcriptome (8, 9).

The development of ribosome profiling (10, 11), a deep sequencing technique in which the ribosome-protected mRNA fragments are sequenced resulting in a genome-wide base resolution signal of translation, allowed this approach to be pushed even further. With ribosome profiling, a proteogenomic search space can be extracted from the translatome, one extra step closer to the final protein products (6, 12–14). In this way, proteogenomic search spaces got even more

---

From the ¹BioBix, Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modeling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium; ²OHMX.bio, Ghent, Belgium; ³Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany; ⁴Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium; ⁵VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
*For correspondence: Gerben Menschaert, Gerben.Menschaert@UGent.be.

attuned and smaller in size because ribosome profiling un-covers both the reading frame of the translated protein product as well as its position inside the transcript. Recently, we could demonstrate that this interplay of ribosome profiling and proteomics allows the validation of novel proteoforms (15) with evidence both on the levels of translation and peptide detection. This approach was embedded in a computational pipeline, called PROTEOFORMER (16).

Despite the fact that these more tailored approaches limit the search space expansion, additional sequences are still added on top of the reference information. So, in order to counteract the effect of this enlarged database, a pursuit for more stringent peptide-to-spectrum matching is ever present in the proteogenomic field. Besides, more stringent matching is always desired as it results in more confident proteomic identifications (2).

The application of postprocessing tools on search engine scores and additional PSM-derived features provides a way to raise the number of spectrum identifications at a fixed FDR (17). Machine learning algorithms can be implemented to separate true from false PSMs using the features obtained from the prior matching step. This concept got especially popular with the introduction of Percolator (18, 19), a tandem mass spectrometry (MS/MS) postprocessing tool that applies support vector machines in combination with a statistical iterative framework to enhance the information yield obtained from PSM features. As such, the confidence to distinguish true from false PSMs can be enhanced.

Beside postprocessing tools, another trend in proteomics is the rise of fragmentation spectrum intensity predictors. Conventional MS/MS scoring engines only consider m/z values when comparing experimental spectra with theoretical ones (20, 21). However, it has been demonstrated that identification rates firmly rise when fragment intensities are added to the search algorithms (22–24). In the last few years, this field has seen a significant progress where machine learning algorithms allow increasingly better predictions of spectral intensity patterns (25–29). One of these predictors is called MS^{2}PIP, and it is trained using gradient decision tree boosting approaches (25, 26). MS^{2}ReScore, a tool linked to MS^{2}PIP, uses MS^{2}PIP's predictions to come up with an extensive feature set, which can then be used to increase the power in PSM post-processing (30). On the other hand, Prosit is a tool that strives to optimally predict peptide spectra through the training of its deep learning model on the comprehensive ProteomeTools database (29). ProteomeTools offers an extremely broad resource of synthetically generated peptides linked to their MS/MS analysis results (31). Both MS^{2}Rescore and Prosit thus offer an extended feature set that is compatible with Percolator postprocessing (29, 30). The combination of spectral intensity predictors with postprocessing tools already demonstrated its advantages for metaproteomics (29), another field that, just like proteogenomics, suffers from FDR problems because of search space size explosion.

In this paper, we provide the first results that this approach is extremely useful in the context of proteogenomics. We demonstrate that this setup clearly improves our earlier ribosome profiling-driven proteogenomics research, leading toward more confident peptide identifications and subsequently more confident proteoform validations. Results show that this leads to considerably more identifications with higher confidence. Furthermore, we also returned to an RNA-Seq–based three-frame translation database and demonstrate that our approach even allows to perform confident peptide identification in this setting, where the search space is around 20 times larger than a reference proteome search space. This last setup moreover shows that third-generation cDNA and direct RNA-Seq using nanopores (32) can now be easily integrated into proteogenomic research.

EXPERIMENTAL PROCEDURES

Construction of Proteogenomic Search Spaces

All experiments were performed on the human colorectal cancer cell line HCT116. Ribosome profiling data from the following paper (6) was used and is available in the Gene Expression Omnibus (dataset GSE58207). In our previous study (16), we applied the PROTEO-FORMER pipeline to derive a search space in FASTA format out of this dataset. The obtained database is a combination of novel candidate translation products derived from ribosome profiling as well as canonical sequences from UniProt.

Next, we performed RNA extraction and purification on HCT116 cells, conversion to cDNA, library preparation, and subsequent Oxford Nanopore Technologies cDNA sequencing on HCT116 cells. Full details on RNA purification, library preparation, and sequencing protocols can be found in the supplemental Materials and supplemental Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41) of this article. Raw sequencing data were uploaded on NCBI's Sequencing Read Archive (Project number SRP289438). Raw sequencing data were base called and mapped. Afterward, translated transcripts were called, and in these, open reading frames (ORFs) were searched over the three reading frames. All ORFs were exported in FASTA format and merged with reference information from UniProt. More details on this data processing can be found in the supplemental Materials and supplemental Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41) of this paper.

MS/MS Data and MaxQuant Search Results

Raw MS/MS data were generated for HCT116 cells in our previous study (16). The experimental protocols applied to obtain the proteomics raw data can be read in the supplemental Materials and supplemental Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41) of the respective publication. The raw data itself are available from ProteomeXchange under the identifier PXD011353.

As the ribosome profiling-based search space is used from that same study (16), the MaxQuant search results (folder: HCT116_uniprot_canonical_txt.zip) under identifier PXD011353 could be used as a starting point for the postprocessing analysis based on the ribosome profiling search space.

For the RNA-Seq case, we searched the same HCT116 raw spectral data against the newly generated RNA-Seq–based ORF FASTA database (merged with the splice isoform-included version of UniProt). Apart from the search space, all search parameters were the same as described in the supplemental Materials and supplemental

Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41) of our previous publication (16). All new search results were submitted to ProteomeXchange under the identifier PXD022280. The MaxQuant search results on protein group level of this search are also available in Excel format in supplemental File S1.

Spectral Intensity Prediction and Postprocessing

For the ribosome profiling case, both MS²ReScore and Prosit were run on the MaxQuant PSM search results. Both tools internally predict the MS/MS spectra (including fragment intensities) for all possible peptides and afterward construct additional features per PSM based on the spectrum prediction of the matching peptide and the experimental spectrum. As these features are combined with the canonical scores of MaxQuant (which do not contain any fragment intensity information, only information from the m/z dimension), the feature information content per PSM increases vastly. Further details on how both tools were used can be found in the supplemental Materials and supplemental Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41). Each of both spectrum predictors puts some constraints on the peptides for which a spectrum can be predicted. These constraints are, for example, based on peptide sequence length, primary fragment ion charge, peptide modifications, amino acid usage, and secondary matches. Further details on these constraints can be found in the supplemental Materials and supplemental Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41). We filtered the total pool of PSMs for a set of constraints in order to obtain a common pool of spectra that can be predicted by both tools. For other analyses in this article, where only one of both spectral predictors was used, this PSM filtering was not performed and all PSMs that could be rescored by that tool were consequently used. Percolator (version 3.02.1) (18, 19) was subsequently used to rescore the common pool of PSMs for different feature combinations: (1) a baseline setting with only the Andromeda score and delta score from MaxQuant, (2) the baseline setting combined with the features from MS²Rescore, and (3) the baseline setting combined with the features from Prosit. For all Percolator runs, the target-decoy competition method was used to obtain q-values. Furthermore, the option to not remove redundant peptides (keep all PSMs) was selected. All other Percolator parameters were set to the default values.

To investigate the feasibility of our approach on a full RNA-Seq–based three-frame translation search space, MS²ReScore was run on the MaxQuant search results for the RNA-Seq protein sequence database with the same settings as described earlier. All rescored PSMs were postprocessed by Percolator for two distinct feature sets: (1) a baseline setting with only scores based on MaxQuant and (2) the baseline setting combined with all features from MS²Rescore. Percolator parameters were set as earlier described.

We also investigated the yield of our new approach on the protein level (next to the PSM and peptide levels), as this can provide information on novel proteoforms, *i.e.*, proteogenomic novel events. Therefore, for the ribosome profiling setting, we linked all PSMs (without prior FDR filtering) that were rescored with Prosit with the protein information obtained in MaxQuant. For the RNA-Seq setting, we used all rescoring information from MS²ReScore. To match the complete MaxQuant search space, we added decoy and contaminant protein sequences to the FASTA file as is done internally in MaxQuant. With the extended Prosit PSM feature file and the extended FASTA file, we ran Percolator again. Compared with previous Percolator analyses in this publication, we included the Fido protein inference algorithm this time (42). The identified proteins were compared against the list of proteins we identified earlier with MaxQuant (16). Furthermore, the proteins, inferred by Fido, were further parsed and classified. We checked per peptide if it validated the existence of a proteoform that got into our search space solely because of ribosome profiling information and not because of UniProt reference information. If so, this protein was added to the list of validated novel proteoforms. Moreover, we made sure that the confirming peptide was included in or covering the protein's variation point (*e.g.*, truncation, extension, splice isoform change, single amino acid variation). Based on this verification, the novel proteoform could also be included in a proteoform subcategory based on the nature of its variation point. Results on the level of proteins and proteoforms are easier to interpret in the ribosome profiling setting as this sequencing strategy clearly delineates the actively translated ORF. Therefore, classification of proteoform results into annotation (sub)classes was primarily done for ribosome profiling (more on this in the supplemental Notes).

The hardware specifications, software availability, and the general statistical rationale can be found in the supplemental Materials and supplemental Figs. S12-S16 (16, 18, 19, 25, 26, 29, 30, 33–41).

RESULTS

To start off, custom protein search spaces were constructed for human HCT116 cells out of ribosome profiling and RNA-Seq analyses. For ribosome profiling, the protein search space obtained with PROTEOFORMER 2.0 (16) was used. For RNA-Seq, HCT116 cells were subjected to Oxford Nanopore Technologies cDNA sequencing, and for each transcript with transcription evidence, all theoretical ORFs that were present over the three reading frames were added to the search space. For both sequencing techniques, the result database was combined with the reference database (SwissProt+TrEMBL) from UniProt to make the distinction between known and novel proteins (proteoforms) later on. In both cases, this results in an expansion of the database size (Table 1 and supplemental Fig. S1). The UniProt reference database (containing 71,356 canonical protein sequences; 93,275 when splice isoforms were included) was expanded to 186,627 sequences owing to the added protein candidates from ribosome profiling, which is a growth of the database size by a factor around 2. Based on the amino acid content, this is a growth by a factor around 1.7. On the other hand, added RNA-Seq results led to a combined database size of 4,988,183 sequences, a total growth in database size by a factor of around 53.5. Based on amino acid content, the RNA-Seq database grew with a factor around 20.4 compared with the UniProt reference.

Using MaxQuant (37), the HCT116 proteomics data were searched against both expanded search spaces. The statistics of these searches (Table 1) illustrate how enlarged database sizes introduce FDR issues in conventional proteomic search algorithms. Compared with a UniProt-only database, a database with added ribosome profiling information only led to a minor decrease in identified PSMs and peptides. The number of identified protein groups even increased as this database has a larger information content, which earlier led to the identification of additional proteoforms (16). For RNA-Seq, this situation, however, got out of hand. The search space grew exponentially, and this resulted in a significant decrease of identified PSMs, peptides, and even protein groups. Using conventional search

TABLE 1
MaxQuant identification statistics of searches against several search spaces from differing sources and sizes

Sequencing technique	UniProt	Search space		Identified protein groups	Identified peptides	Identified PSMs	Identified PSMs
		Entries	Amino acids	MaxQuant	MaxQuant	MaxQuant	MaxQuant+Percolator
None	Canonical	71,356	24,055,511	4294	28,443	180,526	186,937
Ribosome profiling	Canonical	176,202	40,603,175	4333	28,402	177,473	185,767
	Spliced	186,627	46,830,033	4347	28,372	176,978	184,578
RNA-Seq	Spliced	4,988,183	757,075,232	3669	15,820	91,232	175,775

The size of the search space is given based on the number of present sequences as well as based on amino acid content. Information of both ribosome profiling and RNA-Seq could be combined with reference information from UniProt (only canonical proteins or with additional splicing isoforms included). The obtained proteogenomic search spaces were afterward used in the MaxQuant search tool. The number of identified PSMs, peptides, and inferred protein groups clearly differ based on the size of the used search space. Especially for the RNA-Seq–based search space, the size of the search space has dramatic effects on the identification in MaxQuant. Percolator helps to overcome already a big part of this identification reduction. "MaxQuant+Percolator" is used in the rest of the article as the baseline.

algorithms, the extra information added from RNA-Seq got completely undermined by the problems introduced by the search space size expansion. Percolator postprocessing on the raw MaxQuant scores (*i.e.*, Andromeda scores) allowed recovery of the identification rate almost completely for ribosome profiling-based databases, though. For RNA-Seq databases on the other hand, a fair loss of identification seems still present. The "Andromeda+Percolator" setting is in the rest of this article used as the baseline setting.

Based on supplemental Fig. S2, it appears that Max-Quant allows identification of novel proteoforms (*i.e.*, their sequence originally added to the database because of sequencing information) for both the ribosome profiling– and the RNA-Seq–based database. However, the number of identifications supported by both the UniProt reference and the custom sequencing information drops dramatically in the RNA-Seq case. As this is a consequence of the increased search space size and the linked FDR difficulties, improvements to counteract this should first be implemented before one can trustfully validate novel protein identifications. Furthermore, these improvements could also be helpful for the ribosome profiling case as it will help to validate new proteoforms even more confidently.

With that goal in mind, the spectrum predictors MS[2]PIP and Prosit were applied to construct predicted spectra for all peptides in the merged databases. In conventional search tools such as Andromeda (21) (embedded in the MaxQuant (37) software), theoretical spectra are only constructed and matched with experimental spectra based on their x-axis (*m/z* range). Spectral predictors allow to additionally predict the intensities of all fragments of each theoretical spectrum (supplemental Fig. S3), greatly expanding the information content of each PSM. This additional information can be extracted from the PSMs in the form of extra features that can be added to the basic scores from Andromeda. For MS[2]PIP predictions, feature calculation is done using the standalone MS[2]ReScore tool, whereas for Prosit, this is implemented inside the tool itself. Prior to Percolator postprocessing, scores and features could be examined by plotting their distributions for target and decoy PSMs (supplemental Fig. S4). In these plots, it can be observed that the decoy distribution coincides with the lower part of the bimodal target distribution. The targets typically behave bimodal because of the underlying composition of negative and positive PSMs. However, these negative and positive target PSMs are more clearly separated for the predictor scores (panels C and D) than for the conventional scores (panels A and B). For these last scores, there is actually quite some overlap between the negative and positive peak areas, and this effect got even more pronounced for the RNA-Seq case (panels E and F). On the other hand, the principal score of MS[2]ReScore still keeps a clean separation between negatives and positives for the

RNA-Seq case (panel G). The same scores can also be visualized using joint plots (supplemental Fig. S5). In these plots, it is noticeable that the scores of the spectrum predictors allow a cleaner separation than the Andromeda score. This is true in the ribosome profiling case (panels A and B), but it is even more accentuated in the RNA-Seq case (panel E) where there is an even greater mixing on the Andromeda score axis while the MS²Rescore Pearson correlation retains a clean separation between positive targets and decoys.

Concerning scoring, Prosit's spectral angle seems an easier to use statistic compared with MS²ReScore's Pearson correlation as the separation between targets and decoys occurs more spread over the [0,1]-range for spectral angle than for Pearson correlation (supplemental Fig. S5C). An eventual threshold for spectral angle would be positioned more in the center of this range, making separation easier. MS²ReScore also provides a cosine score in its feature vector. This score resembles Prosit's spectral angle more than the Pearson correlation, although it is not completely identical. A comparison between these two scores (supplemental Fig. S5D) shows that both tools succeed very well in separating positive targets from negative targets and decoys.

Next, feature vectors from (1) solely Andromeda, (2) Andromeda combined with MS²ReScore features, and (3) Andromeda combined with Prosit features were, for the ribosome profiling case, processed with Percolator to check whether the features added by the intensity-based predictors boosted the identification process. For the analysis starting from the RNA-Seq search space, only the first two vectors were compared. In its postprocessing, Percolator generates total scores and statistical measures per PSM (supplemental Files S2–S4 for the ribosome profiling case; supplemental Files S5 and S6 for the RNA-Seq case). Distributions of these scores give a first impression of the identification performance of Percolator for the different feature sets (supplemental Fig. S6). For a feature set with only Andromeda scores, there is quite some distribution overlap between the true and false target PSMs, especially for the analysis started from the RNA-Seq search space. Addition of intensity-based features makes the peak areas of these bimodal distributions sharper as these added features allow Percolator to strengthen its separating power. Also, the confidence on which individual PSMs can be identified rises when spectral prediction information is added, as there is enrichment for lower posterior error probabilities (PEPs) visible for feature sets with additional intensity-based features included (supplemental Fig. S7).

Filtering for a set of significant PSMs is generally performed using the q-values and the resulting FDR-estimation (18). The most valuable measure to compare the sets with and without additional features is therefore the FDR as this shows the actual yield on PSM-level stringency. In Figure 1, it is clearly visible that the introduced MS²ReScore and Prosit features extensively improve the stringency on which Percolator can classify the PSMs. This is illustrated by the remarkable shift to the left, compared with the baseline. Next to that, there is also a gain in identification rate, which is shown in the vertical direction of this figure. Moreover, for our data, it appears that MS²ReScore and Prosit perform this task almost equally well. The number of identified spectra could also be converted to the number of true-positive spectra using following formula:

$$\#positives = \#identifications - FDR \cdot \#identifications \quad (1)$$

The true positives plot is given in supplemental Fig. S8, and also here, the classification stringency of Percolator is shown to improve. Using additional features from spectral predictors leads to both a loss and gain of PSMs compared with conventional search strategies (supplemental Fig. S9). The loss describes the PSMs that appear to be false matches after all, while the gain describes PSMs that match much better as the fragment intensities are taken into account. In supplemental Fig. S9, it is shown that the gain of using extended features is bigger than the loss for FDR thresholds up until 0.1% and sometimes even lower, compared with a conventional setting of using only Andromeda features at a 1% FDR threshold.

In the context of proteogenomics, it is interesting to push the improved PSM information onto the protein level. This allows validation of the translation profiles of ORFs from ribosome profiling or the transcript expression from RNA-Seq with protein validation from proteomics. We performed this using conventional proteomic search strategies before for ribosome profiling databases (6, 12, 13, 16), but incorporating the improved PSM information into this workflow would provide significant advantages. In order to demonstrate this, we used the extended feature set of Andromeda combined with Prosit for PSMs coming from a search against the ribosome profiling custom search space. This set was again inputted in Percolator but this time with the Fido (42) protein inference algorithm activated. The Percolator results (supplemental Files S7–S9) were further parsed to search for novel proteoforms, similar to what was done before based on MaxQuant output (16). In doing so, the confidence on which proteoform identifications could be called was checked. The novel proteoforms obtained from this novel Prosit-Percolator approach are given in supplemental File S10. The collections of novel proteoforms are compared between the novel Prosit-Percolator approach and the conventional approach using MaxQuant (16) in Figure 2. Panel A shows that a big part of the novel proteoforms is shared between the two approaches, but nevertheless, both analyses identify unique novel proteoforms as well. Further investigation into these unique proteoforms for each technique shows, however, that the underlying peptides of the Prosit-Percolator approach are distributed over an expected confident PEP (posterior error probability) range (panel C),
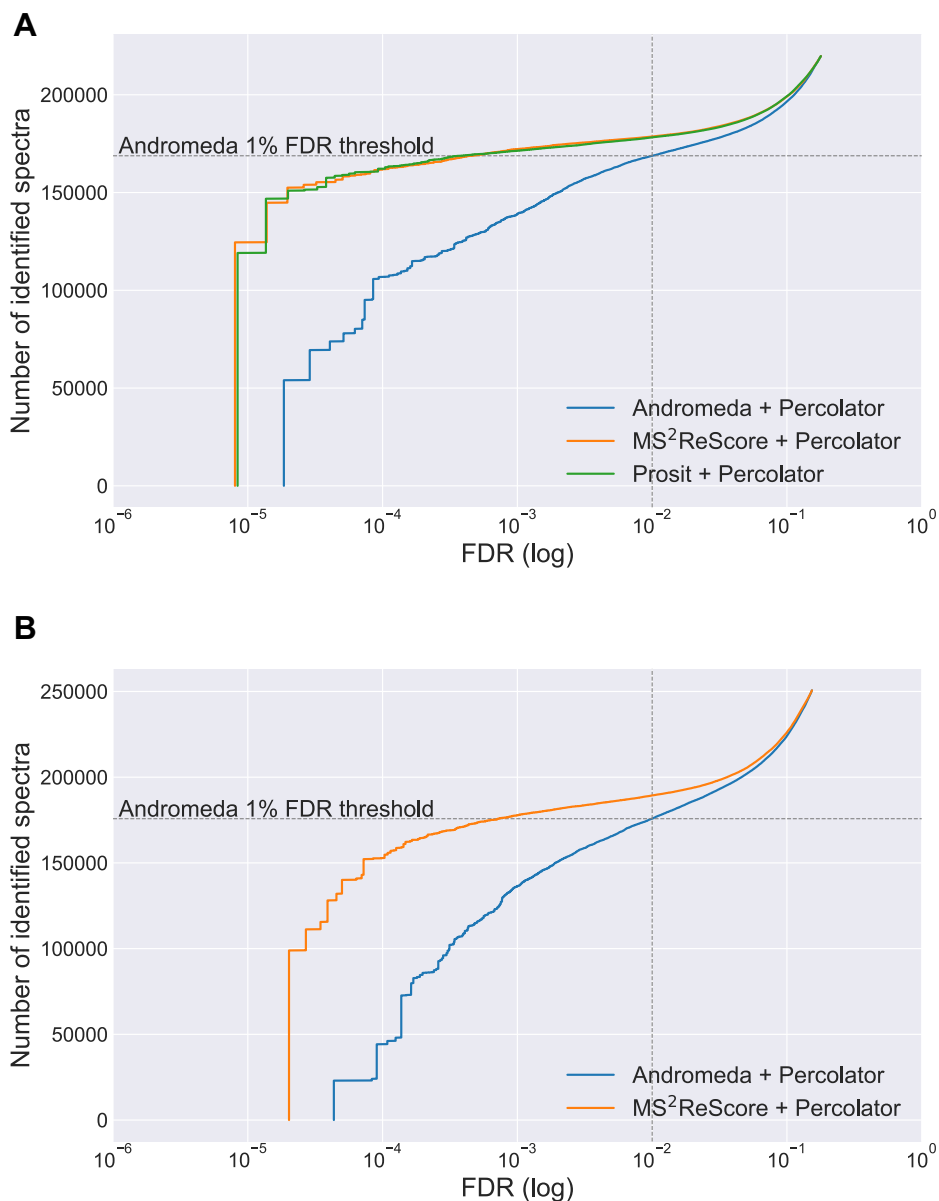
Fig. 1. **The number of identified spectra as a function of differing FDR levels for the different feature sets.** *A*, for the ribosome profiling case, the feature set of only Andromeda scores (serving as a baseline), the combination of Andromeda scores with MS^{2}ReScore features, and the combination of Andromeda scores with Prosit features are shown. *B*, for the RNA-Seq case, the feature set of only Andromeda scores (serving as a baseline) is compared with the combination of Andromeda scores and MS^{2}ReScore features. Percolator was used on both baseline and expanded feature sets. Additional intensity-based features on top of the Andromeda scores enhance Percolator's capability to separate true from false peptide-to-spectrum matches, ultimately leading to both a higher identification rate as well as an elevated stringency. For example, if one would use Andromeda + Percolator at a 1% FDR threshold, extended features will allow one to identify around 10,000 PSMs more (described by the *vertical dotted line*). At the same time, these identifications will be much more stringent and confident as the underlying confidence measures of these individual identifications will be much smaller (described by the shift to the *left*). FDR, false discovery rate.

whereas the underlying peptides of the MaxQuant approach are situated around the FDR threshold of 1% (panel B). The MaxQuant-unique peptides thus rather not present the most confident identifications and are classified as doubtful by the Prosit-Percolator approach. The increase in identification rate is therefore present on two levels: in absolute numbers of identified proteins as well as by replacing doubtful MaxQuant identifications with more confident ones using Prosit-Percolator. For the shared proteoforms (panels D-F), the PEP distribution of the underlying peptides is generally situated at more stringent values for the Prosit-Percolator approach than for the MaxQuant, further illustrating that also on the protein level the additional PSM information leads at the same time to more confidence.
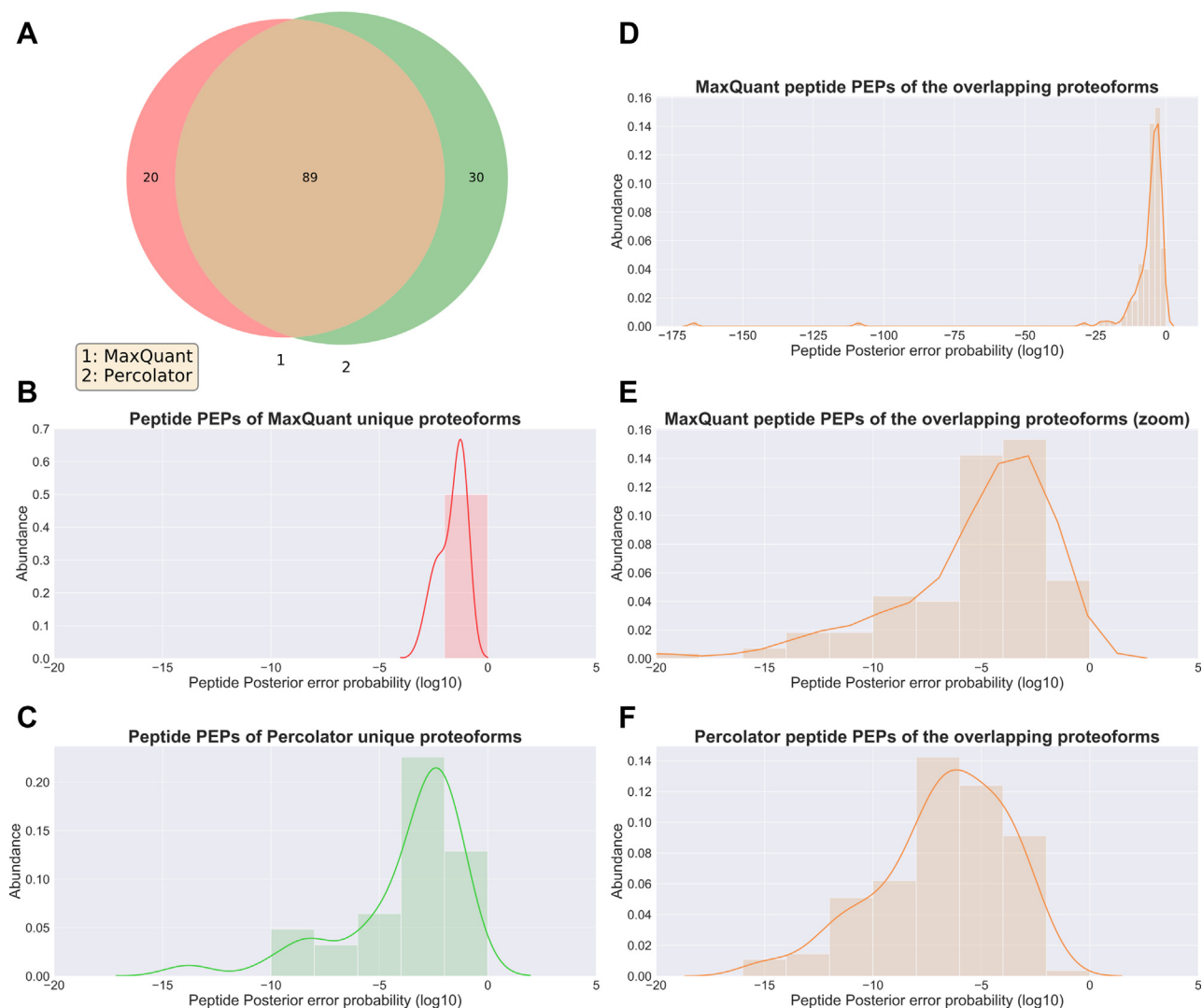
FIG. 2. Comparison of the novel proteoform identification results (ribosome profiling setting). The new approach, using Prosit and Andromeda features in Percolator rescoring, was compared with the conventional approach, using direct Andromeda results in MaxQuant (without rescoring). *A*, Venn diagram showing the overlap in novel proteoforms found by both approaches. *B*, PEP scores of the peptides confirming novel proteoforms, unique for the MaxQuant approach. *C*, PEP scores of the peptides confirming novel proteoforms, unique for the Prosit-Percolator approach. *D*, PEP scores, as obtained from the MaxQuant approach, of the peptides for the shared novel proteoforms. Two outliers were situated at the *left side* of the *x*-axis. *E*, analog to (*D*), but zoomed in on a specific range of the *x*-axis. *F*, PEP scores, as obtained from the Prosit-Percolator approach, of the peptides for the shared novel proteoforms. PEP, posterior error probability.

As done in our previous work for the conventional Max-Quant approach ([16]), it is interesting to classify the novel proteoforms based on the nature of their variation point. However, this time, the proteomics results with the intensity-based features were included. In Figure 3, the classification of all novel proteoforms found with the new Prosit-Percolator approach is presented. Around one-third of the novel proteoforms occurs because of splicing events, whereas another third because of translation in generally presumed noncoding regions. Furthermore, events like N-terminal extensions, N-terminal truncations, out-of-frame ORFs, upstream ORFs, downstream ORFs, and single amino acid variations could be validated with proteomics. In order to compare the novel

Prosit-Percolator approach with the conventional MaxQuant approach, classification results were put next to each other in supplemental Fig. S10. Remarkable classification result differences between these two approaches are elaborated on in the discussion section below.

Proteoform identification results were also analyzed for the RNA-Seq setting (supplemental Fig. S11). The proteoform overlap between MS[2]ReScore-Percolator and MaxQuant is now smaller, though, which is expected as MaxQuant has more problems because of the even larger database of RNA-Seq. Therefore, MaxQuant misses a lot of the novel proteoforms and, at the same time, MaxQuant also calls some doubtful proteoforms, as can be seen based on the PEP distribution
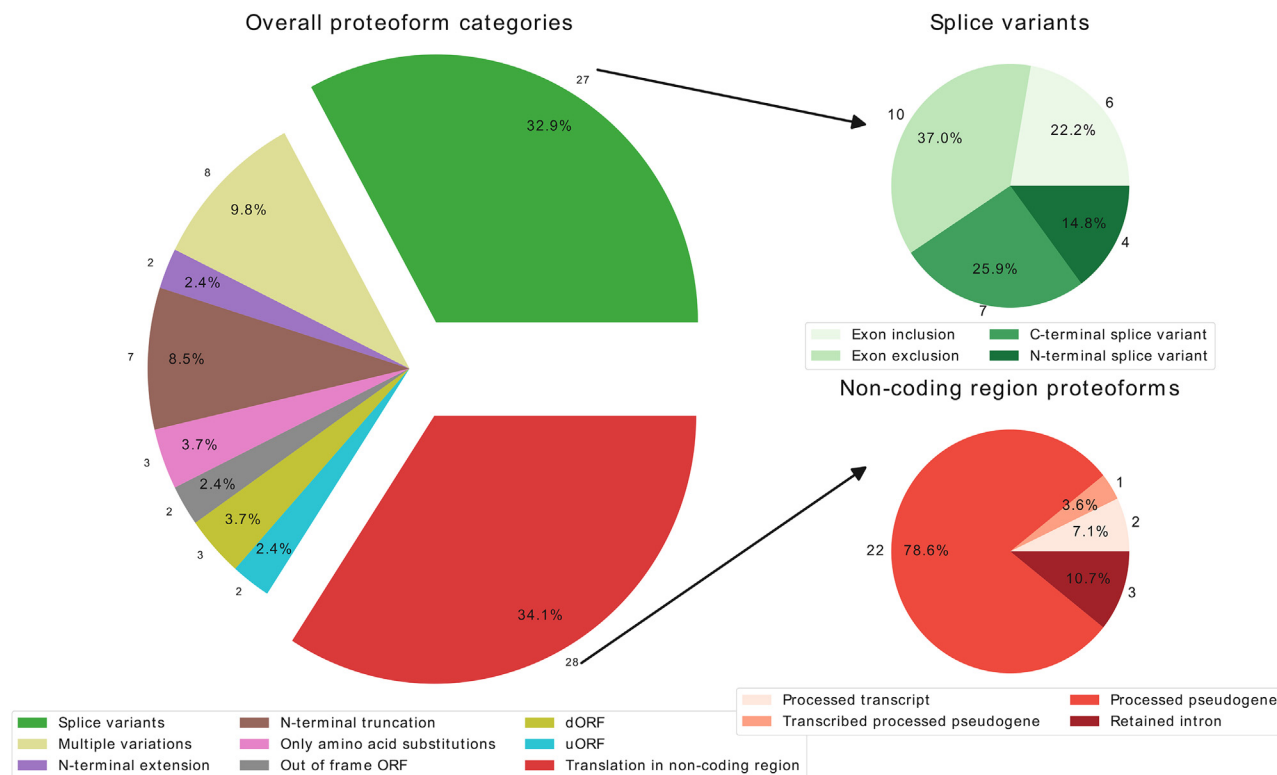
Fɪɢ. 3. **Classification of the MS-validated proteoforms found with Percolator using extended Prosit features based on the ribosome profiling search space.** The proteoforms are classified based on the nature of their variation. For splice variants and proteoforms in previously considered noncoding regions, more detailed classification categories are shown. More details about the different classification categories can be found in supplemental Table S1.

(Panel B). The PEP distributions of MS^{2}ReScore-Percolator, however, remain unaffected by the database size (Panel C). Also, for the PEP distributions of the novel proteoforms that the two approaches have in common, a negative effect of the database size on the MaxQuant PEPs can be observed (Panel D), whereas this effect is absent for MS^{2}Rescore-Percolator (Panel E). Taken together, the differences in the proteoform identification rate between the RNA-Seq setting and the ribosome profiling setting can be accounted onto following two causes. On the one hand, the RNA-Seq database is too vast for MaxQuant to call all present novel proteoforms correctly. On the other hand, the RNA-Seq database contains even more search space information than the ribosome profiling database, leading to extra identifications with the MS^{2}Rescore-Percolator approach, as this new algorithm allows one to circumvent the issues typically seen for comprehensive search spaces.

DISCUSSION

Over the last years, machine learning algorithms transformed mass spectrometry–based proteomics research ([18], [25], [28], [29]). Conventional database search algorithms largely base their scoring on the presence of specific fragment ions but mostly ignore the signal intensities of these fragment ions

([20], [21], [43–45]). However, it was demonstrated that the introduction of *in silico* predicted intensity information into the search strategy improves the peptide identification rates significantly ([22–24]). To use the full information of fragment intensities, it was shown that the predictions of MS^{2}PIP and Prosit can be used to rescore PSMs in order to improve confidence and peptide identification rates in standard proteomic ([29], [30]). However, this was not yet applied in proteogenomics, a field that specifically encounters problems with decreased confidence and peptide identification rates because of enlarged search space sizes ([1]).

In order to demonstrate the applicability and the advantages of these new machine learning–based PSM rescoring techniques in proteogenomics, search spaces were constructed that merge reference information from UniProt with novel candidate protein information from sequencing techniques. We previously demonstrated how ribosome profiling can help in tempering this search space expansion while still providing additional information from the translatome layer. In this way, ribosome profiling allowed outlining enough novel proteoform candidates without creating an overload of useless sequences. As such, proteoform candidates can still be validated confidently with conventional proteomic search strategies afterward ([16]). For a first application of the novel

rescoring techniques in proteogenomics, this ribosome profiling–based search space provides a good opportunity to check the yield on a moderately expanded search space. In addition, a proteogenomic search space was also constructed based on a nanopore RNA-Seq run. This resulted in a database that is around 15 times bigger (based on amino acid content) than the database obtained from ribosome profiling (Table 1 and supplemental Fig. S1). In contrast to ribosome profiling, RNA-Seq does not account which transcripts are subjected to translation, neither does it allow selection of the part of the transcript that leads to a protein. On top of that, ORFs generated over three reading frames are presented in the RNA-Seq database, while the triplet periodic signal in ribosome profiling allows one to select for ORFs in the active reading frame (46). Therefore, RNA-Seq databases contain all possible ORFs that occur in every transcribed transcript and a lot of these are in fact not translationally active. As we did not apply additional filtering on the sequences and as all transcripts with at least one read were retained, very lenient thresholds were here applied on purpose. This results in an even bigger search space as compared with previous RNA-Seq–based proteogenomic studies (8, 38, 47–49), but as the goal was here to test to which degree machine learning–based rescoring techniques can compensate this effect, such an expansion rate was actually what was being pursued.

In order to increase the peptide identification stringency, machine learning–based rescoring techniques apply matching scores that take into account the fragment intensity information of both spectrum predictions and experimental spectra (29, 30). This information was often overlooked in conventional matching strategies (21). Overfitting is very unlikely to occur in this rescoring process as the training sets of the applied spectrum predictors (25, 26, 50) are completely unrelated to the data used in this study. Furthermore, both spectrum predictors apply different regularization techniques (cross-validation, early stopping (51), dropout (52), inherently noisy data) to avoid overfitting (25, 26, 29). Percolator, which combines these features and tries to maximize the discriminative power between true and false PSMs, is also insusceptible for overfitting as it uses a cross-validation strategy and, in addition, because it exploits different features without focusing on certain types of spectra (18). On top of that, using Percolator in combination with one of both rescoring tools does not contribute to additional overfitting as the two tools function as successive steps that optimize different problems.

As shown in supplemental Figs. S4 and S5, the advantages of incorporating machine learning–based scores are already visible in the distributions of these scores. Compared with canonical scores like the Andromeda score, these novel intensity-based scores, such as Pearson correlation and spectral angle, draw upon a broader information content as they use not only the m/z values of fragment ions but also their intensities. As such, another dimension (y-axis

representing the intensity) is added on top of the already included m/z value (x-axis of the fragmentation spectrum). This is reflected in the clearer separation of decoys and positive target PSMs for the novel features. In Percolator, the total information that can be learned from both novel and conventional scores is maximized, and as almost literally illustrated in the joint plots of supplemental Fig. S5, a dimension of yet overlooked PSM information is as such included.

Percolator on itself is already a powerful tool, which is reflected in the data presented here. For example, in Figure 1B and Table 1, Percolator identifies 175,775 PSMs for the RNA-Seq search space based on a feature set that only contains conventional scores from Andromeda using the default 1% FDR threshold. MaxQuant also uses these same Andromeda scores (37) but hardly identifies 91,232 PSMs for the RNA-Seq database (Table 1). On top of that, it can be seen that Percolator manages to do even a far better job when the more informative intensity-based features are also added to the feature set. This is visible in the distributions of the Percolator scores (supplemental Fig. S6), where intensity-based features allow a cleaner separation of positives and negatives, and in the PEP distributions (supplemental Fig. S7), where the additional features enrich more confident values. Furthermore, and important toward better identification results, these extended feature sets allow one to filter more stringently and with higher PSM identification rates, as shown in the FDR and true-positive plots (Fig. 1 and supplemental Fig. S8). It is even the case that some false identifications (initially found in the baseline search) get replaced with identifications that are more plausible in the new step because extra information is added in the postprocessing (supplemental Fig. S9). Taken together, the advantages of the novel approach are thus dual. On the one hand, features from spectral predictors elevate the identification rate. On the other hand, the identification stringency greatly improves as the positive identifications can be called with much more certainty (Fig. 1). This dual advantage is also illustrated on the level of novel proteoforms in Figure 2, where panels B and C describe the gain of additional identifications, whereas panels E and F rather describe the gain in stringency.

Performing PSM identification more stringently is definitely a positive thing because as such, one can expect more identifications to be truly positive. Moreover, there is always a chance that a correct PSM is ranked lower than a false PSM for a certain experimental spectrum because of a lack of information when only m/z values are taken into account. This is certainly the case for larger search spaces as the more theoretical candidates are available, the higher the likelihood that the best match for a specific experimental spectrum is an incorrect one by random chance (1). Including knowledge about intensities will thus help the correct PSM in jumping over the false-positive PSMs in the ranking, making

it more likely that it will be the best scoring PSM for that spectrum. It should be noted, though, that this jumping in the ranking of PSMs per experimental spectrum is not implemented in our algorithm yet. In this first application on proteogenomics, we only rescored the best scoring PSM per experimental spectrum. Rescoring multiple PSMs per spectrum would require one to thoroughly check how the decoy and target distributions behave as the balance between the number of positive and negative PSMs in the total distribution would shift completely when multiple matches per spectrum are allowed. Nevertheless, it is an interesting future implementation as it could prove to be useful for the proteogenomic identification of additional novel proteoforms. Other previous efforts have tried to incorporate predicted spectra and their intensity information in the proteomic search engine itself (53), rather than using this for rescoring purposes. In that way, intensity information is already available when the best matching PSM for each experimental spectrum is picked. Besides this, intensity predictions could also help in resolving chimeric spectra (54).

Over all the results in this publication, the fairest comparison between Prosit and MS^{2}ReScore can be made based on the FDR plots (Fig. 1*A*), as herein it is given how much the additional features yield in peptide identification rates and stringency. It is clear that both tools perform equally well for the purposes we wanted to compare and demonstrate in this proteogenomic study.

The raw proteomic data that were used here (16) were acquired in quite standard conditions (cell conditions, MS/MS settings, fragmentation method, digestion, etc.). It would be interesting, though, to check how the novel tools perform in different sets of conditions. MS^{2}PIP trains a model for each new condition set (fragmentation method, MS/MS analyzer, labeling technique) (25, 26), whereas Prosit presents its training basis to be more robust to changing conditions (29), mainly owing to its deep learning architecture and its vast training dataset (31, 50). Because of this, it can be expected that Prosit is able to predict better in more versatile conditions without any retraining. On the other side, MS^{2}PIP's model is less complex to use and does not depend on graphical processing units, which makes it applicable to a broader range of computational resources. For the analysis of datasets digested with other enzymes than trypsin, MS^{2}PIP is for the moment no option as it is currently only trained on trypsin data (26). In the future, MS^{2}PIP models for other digestion enzymes will become available (internal communication). Prosit, however, showed to be able to transfer its learning over different digestion enzymes (29).

In previous work (16), we demonstrated how the unique hallmarks of ribosome profiling can be used to obtain a set of novel proteoform candidates, which could subsequently be validated using matching proteomics. The proof-of-concept studies presented here show that machine learning–based approaches can improve this proteomic validation of proteoform candidates tremendously. This was true not only on the

level of PSMs and peptides (Fig. 1) but also on the protein level (Fig. 2) as less good MaxQuant matches are removed (Fig. 2*B*) and novel truthful identifications get over the filtering threshold. On the protein level, Percolator on itself again seems to have a share in this process. The additional features of spectral predictors further increase the discriminative power on the protein level as well.

With enhanced power and stringency, we are thus now able to validate novel proteoforms on the proteomic level. These novel proteoforms can be subdivided in different categories, based on the nature of their variation, as shown in Figure 3. The abundances of these different proteoform categories in our novel Prosit-Percolator approach can be compared with the abundances coming out of the conventional approach using MaxQuant (16). This comparison is given in supplemental Fig. S10. A few findings catch the eye here. First, the number of proteoforms with N-terminal variation points (N-terminal truncations, extensions, and N-terminal splice variants) was higher using MaxQuant. This has a logical explanation, as at the time of our analysis, the latest version of Prosit could not take into account peptides with N-terminal acetylations. Of course, N-terminal variants are mostly supported by N-terminal peptides and these generally contain an acetylation at their N terminus. Second, the "multiple variations" category decreases using Prosit-Percolator. In the prior MaxQuant results, it is more likely that one of the variation points of a proteoform with multiple variations is actually false than that the variation point of a proteoform with only one variation would be false. Therefore, using the more stringent Prosit-Percolator approach will rather eliminate false-positive proteoforms from the category with multiple variations. And third, the novel Prosit-Percolator approach results in more novel proteoforms from earlier supposed noncoding regions, especially from pseudogenes. It is known that pseudogenes could, especially in specific cell conditions or differentiation states, still be expressed (55, 56). Owing to the increased confidence of using extended PSM feature sets, the novel approach will be stronger in finding proteomic evidence for eventual translation events in pseudogenic regions. On top of that, this analysis is done in a specific differentiation and disease state, namely, a colon cancer cell line, which could explain extra translation in pseudogenic regions. Besides that, analyses that used prefractionation techniques in order to improve proteomic identification stringency also identified an enrichment of translation evidence in pseudogenes in the proteomics search results (57). Enhanced search strategies altogether could thus tend to enrich translation evidence from pseudogenic regions in a similar way.

Next to analyses on ribosome profiling–based custom search spaces, we tested our approach also in RNA-Seq–based search spaces that are several times larger in size. With conventional proteomics, this database size expansion would imply severe problems for statistical validation, but we could

demonstrate that most of these issues can be compensated for by using spectral prediction features in combination with Percolator. This makes the urge for custom search spaces with manageable sizes less pressing. With these new tools in place, statistically feasible proteogenomics could thus be possible, not only with custom search spaces coming from ribosome profiling, but also with search spaces from conventional RNA-Seq analyses. This has some practical advantages. Ribosome profiling requires a time- and skill-intensive wet laboratory protocol (11, 46), and specialized software is generally necessary to handle its valuable but specific data hallmarks (58). RNA-Seq on the other hand has a much easier library preparation protocol and is more accessible and more routinely applicable for most laboratories. Also, the computational pipeline for RNA-Seq contains in general less steps and more tools are available as compared with ribosome profiling (59). Therefore, these state-of-the-art proteomic search strategies can make RNA-Seq–based proteogenomics possible and, as such, they can help in making proteogenomics more accessible as an overall approach to more laboratories. This could mean that, in the upcoming years, increasingly more studies will decide to replace general proteomic reference search spaces (*e.g.*, UniProt) with sample-specific custom search spaces based on sequencing information. Of course, this could have significant effects on the outcome of future studies as the reference dataset is said to be far from complete (60–63).

Using RNA-Seq instead of ribosome profiling as the basis for proteogenomics has also another big advantage. Ribosome profiling is oriented toward the Illumina high-throughput sequencing platform, mainly because of its extensive sequencing depth, which allows detection of the ribosomal profile in more detail (11). However, over the last years, the rise and continuous further development of third-generation sequencing techniques was observed (64). Third-generation techniques as SMRT sequencing (65, 66) and nanopore sequencing (67) pushed the upper read length boundary to a whole new dimension. This switch to longer read lengths has major advantages for detecting different transcript isoforms in proteogenomics. Also, it allows an easier characterization of novel splicing events and repetitive regions. Ribosome profiling is not in an urge to make the step toward third-generation sequencing, though, as the ribosome protected fragments are anyway around 28 base pairs in length (68), short enough to be comfortably measured with Illumina sequencing. Furthermore, the third generation is just recently approaching the read accuracy levels of Illumina sequencing (69). However, third-generation sequencing offers additional advantages that could be beneficial in proteogenomic workflows. As such, third-generation sequencing techniques do not require an amplification step before sequencing, which avoids the presence of PCR artefacts in the readout (32). Second, great efforts are invested in minimal sample preparation, portable instrumentation, and lower library preparation times and costs, especially

for the nanopore sequencing technology (70, 71). And third, third-generation sequencing allows the direct study of modified bases in both DNA and RNA (32). As RNA-Seq–based proteogenomics can readily apply third-generation sequencing, as shown in this study, all these additional advantages could be built in into future proteogenomic workflows.

To conclude, we strongly believe that spectral predictors and machine learning–based approaches will play a major future role in computational proteogenomics. As demonstrated in this study, these algorithms come to good use, in particular to counter the statistical drawback of using oversized custom search spaces constructed out of matching sequencing experiments. As these techniques find a more common place in proteogenomics research, we believe that ever more conclusions from the sequencing level can find validation at the proteomic level, further nurturing the cross-fertilization between these two important omics fields.

DATA AVAILABILITY

Raw ribosome profiling reads used in this article can be found in the Gene Expression Omnibus (dataset GSE58207). Raw nanopore cDNA sequencing reads were submitted to NCBI's Sequencing Read Archive (SRP289438). The MS/MS raw spectral data can be found on the website of the ProteomeXchange Consortium (dataset identifier PXD011353). Results of the MaxQuant search against the ribosome profiling–based database can also be found in ProteomeXchange identifier PXD011353. Results of the MaxQuant search against the RNA-seq–based database are deposited under ProteomeXchange identifier PXD022280.

Abbreviations—The abbreviations used are: cDNA, Complementary DNA; FDR, False discovery rate; MS/MS, Tandem mass spectrometry; PEP, Posterior error probability; PSM, Peptide-to-spectrum match.

REFERENCES

1. Nesvizhskii, A. I. (2014) Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125
2. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **73**, 2092–2123
3. Colaert, N., Degroeve, S., Helsens, K., and Martens, L. (2011) Analysis of the resolution limitations of peptide identification algorithms. *J. Proteome Res.* **10**, 5555–5561
4. Blakeley, P., Overton, I. M., and Hubbard, S. J. (2012) Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **11**, 5221–5234
5. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., and Macek, B. (2013) Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* **12**, 3420–3430
6. Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W., Van Damme, P., and Menschaert, G. (2014) PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* **43**, e29
7. Khatun, J., Yu, Y., Wrobel, J. A., Risk, B. A., Gunawardena, H. P., Secrest, A., Spitzer, W. J., Xie, L., Wang, L., Chen, X., and Giddings, M. C. (2013) Whole human genome proteogenomic mapping for ENCODE cell line data: Identifying protein-coding regions. *BMC Genomics* **14**, 141
8. Wang, X., Liu, Q., and Zhang, B. (2014) Leveraging the complementary nature of RNA-seq and shotgun proteomics data. *Proteomics* **14**, 2676–2687
9. Komor, M. A., Pham, T. V., Hiemstra, A. C., Piersma, S. R., Bolijn, A. S., Schelfhorst, T., Diemen, P. M. D., Tijssen, M., Sebra, R. P., Ashby, M., Meijer, G. A., Jimenez, C. R., and Fijneman, R. J. A. (2017) Identification of differentially expressed splice variants by the proteogenomic pipeline. *Mol. Cell. Proteomics* **16**, 1850–1863
10. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223
11. McGlincy, N. J., and Ingolia, N. T. (2017) Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129
12. Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K., and Van Damme, P. (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics* **12**, 1780–1790
13. Koch, A., Gawron, D., Steyaert, S., Ndah, E., Crappé, J., De Keulenaer, S., De Meester, E., Ma, M., Shen, B., Gevaert, K., Van Criekinge, W., Van Damme, P., and Menschaert, G. (2014) A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of
14. Peeters, M., and Menschaert, G. (2020) The hunt for sORFs: A multidisciplinary strategy. *Exp. Cell Res.* **391**, 111923
15. Smith, L. M., and Kelleher, N. L. (2013) Proteoform: A single term describing protein complexity. *Nat. Methods* **10**, 186–187
16. Verbruggen, S., Ndah, E., Van Criekinge, W., Gessulat, S., Kuster, B., Wilhelm, M., Van Damme, P., and Menschaert, G. (2019) PROTEOFORMER 2.0: Further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Mol. Cell. Proteomics* **18**, S126–S140
17. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
18. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925
19. The, M., MacCoss, M. J., Noble, W. S., and Käll, L. (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727
20. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277
21. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805
22. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6**, 654–661
23. Narasimhan, C., Tabb, D. L., VerBerkmoes, N. C., Thompson, M. R., Hettich, R. L., and Uberbacher, E. C. (2005) Maspic: Intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem.* **77**, 7581–7593
24. Sadygov, R., Wohlschlegel, J., Park, S. K., Xu, T., and Yates, J. R. (2006) Central limit theorem as an approximation for intensity-based scoring function. *Anal. Chem.* **78**, 89–95
25. Degroeve, S., and Martens, L. (2013) MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–3203
26. Gabriels, R., Martens, L., and Degroeve, S. (2019) Updated MS^{2}PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.* **47**, W295–W299
27. Zhou, X., Zeng, W., Chi, H., Luo, C., Liu, C., Zhan, J., He, S., and Zhang, Z. (2017) pDeep: Predicting MS/MS spectra of peptides with deep learning. *Anal. Chem.* **89**, 12690–12697
28. Tiwary, S., Levy, R., Gutenbrunner, P., Soto, F. S., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., and Cox, J. (2019) High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat. Methods* **16**, 519–525
29. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H., Aiche, S., Kuster, B., and Wilhelm, M. (2019) Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518
30. Silva, A. S. C., Bouwmeester, R., Martens, L., and Degroeve, S. (2019) Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* **35**, 5243–5248
31. Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., *et al*. (2017) Building proteometools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262
32. Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., *et al* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206
33. Lanfear, R., Schalamun, M., Kainer, D., Wang, W., and Schwessinger, B. (2019) MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics* **35**, 523–525
34. Andrews, S. (2010) FastQC: A quality control tool for high throughput sequence data. *unpublished*

35. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100

36. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419

37. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol* **26**, 1367–1373

38. Woo, S., Cha, S. W., Na, S., Guest, C., Liu, T., Smith, R. D., Rodland, K. D., Payne, S., and Bafna, V. (2014) Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics* **14**, 2719–2730

39. Wang, X., Slebos, R. J. C., Wang, D., Halvey, P. J., David, L., Liebler, D. C., and Zhang, B. (2012) Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **11**, 1009–1017

40. Halvey, P. J., Wang, X., Wang, J., Bhat, A. A., Dhawan, P., Li, M., Zhang, B., Liebler, D. C., and Slebos, R. J. C. (2014) Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. *Cancer Res.* **74**, 387–397

41. Ning, K., and Nesvizhskii, A. I. (2010) The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics* **11**, S14

42. Serang, O., and Maccoss, M. J. (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **9**, 5346–5357

43. Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data proteomics and 2-DE. *Electrophoresis* **20**, 3551–3567

44. Eng, J. K., Mccormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989

45. Craig, R., and Beavis, R. C. (2004) Tandem: Matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467

46. Ingolia, N. T., Brar, G. A., Rouskin, S., Mcgeachy, A. M., and Weissman, J. S. (2012) The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550

47. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., Maccoss, M., and Bafna, V. (2014) Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **13**, 21–28

48. Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., Griffin, T. J., and Smith, L. M. (2014) Using galaxy-P to leverage RNA-seq for the discovery of novel protein variations. *BMC Genomics* **15**, 703

49. Wen, B., Xu, S., Sheynkman, G. M., Feng, Q., Lin, L., Wang, Q., Xu, X., Wang, J., and Liu, S. (2014) sapFinder: An R/bioconductor package for detection of variant peptides in shotgun proteomics experiments. *Bioinformatics* **30**, 3136–3138

50. Zolg, D. P., Wilhelm, M., Schmidt, T., Me, G., Zerweck, J., Knaute, T., Wenschuh, H., Reimer, U., Schnatbaum, K., and Kuster, B. (2018) ProteomeTools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Mol. Cell. Proteomics* **17**, 1850–1863

51. Caruana, R., Lawrence, S., Giles, L., and Giles, R. C. S. L. L. (2001) Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Adv. Neural Inf. Process. Syst.* **13**, 402–408

52. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958

53. Degroeve, S., and Martens, L. (2018) Ionbot: A novel, fully data-driven search engine for open modification and mutation searches with applications in quantitative (meta-)proteomics. *HUPO* **2018**, 42. B. Abstr.

54. Dorfer, V., Maltsev, S., Winkler, S., and Mechtler, K. (2018) CharmeRT: Boosting peptide identifications by chimeric spectra identification and retention time prediction. *J. Proteome Res.* **17**, 2581–2589

55. Chen, X., Wan, L., Wang, W., Xi, W., Yang, A., and Wang, T. (2020) Re-recognition of pseudogenes: From molecular to clinical applications. *Theranostics* **10**, 1479–1499

56. Mei, D., Song, H., Wang, K., Lou, Y., Sun, W., Liu, Z., Ding, X., and Guo, J. (2013) Up-regulation of SUMO1 pseudogene 3 (SUMO1P3) in gastric cancer and its clinical association. *Med. Oncol.* **30**, 709

57. Branca, R. M. M., Orre, L. M., Johansson, H. J., Granholm, V., Huss, M., Pérez-Bercoff, Å., Forshed, J., Käll, L., and Lehtiö, J. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62

58. Kiniry, S. J., Michel, A. M., and Baranov, P. V. (2019) Computational methods for ribosome profiling data analysis. *Wiley Interdiscip. Rev. RNA* **11**, e1577

59. Yang, I. S., and Kim, S. (2015) Analysis of whole transcriptome sequencing data: Workflow and software. *Genomics Inform.* **13**, 119–125

60. Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2015) sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **44**, D324–D329

61. Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018) An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **46**, D497–D502

62. Vanderperre, B., Lucier, J.-F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.-M., and Roucou, X. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8**, e70698

63. Brunet, M. A., Brunelle, M., Lucier, J.-F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J.-D., Dufour, P., Jacques, J.-F., Fournier, I., Ouangraoua, A., Scott, M. S., Boisvert, F.-M., *et al* (2019) OpenProt: A more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **47**, D403–D410

64. Morey, M., Fernández-marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., and Cocho, J. A. (2013) A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* **110**, 3–24

65. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., *et al* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–139

66. Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013) The advantages of SMRT sequencing. *Genome Biol.* **14**, 405

67. Bayley, H. (2015) Nanopore sequencing: From imagination to reality. *Clin. Chem.* **61**, 25–31

68. Verbruggen, S., and Menschaert, G. (2019) mQC: A post-mapping data exploration tool for ribosome profiling. *Comput. Methods Programs Biomed.* **181**, 104806

69. Noakes, M. T., Brinkerhoff, H., Laszlo, A. H., Derrington, I. M., Langford, K. W., Mount, J. W., Bowman, J. L., Baker, K. S., Doering, K. M., Tickman, B. I., and Gundlach, J. H. (2019) Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. *Nat. Biotechnol.* **37**, 651–656

70. Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., Mcintyre, A. B. R., Dworkin, J. P., Lupisella, M. L., Smith, D. J., Botkin, D. J., Stephenson, T. A., Juul, S., Turner, D. J., Federman, S., Stryke, D., *et al* (2017) Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* **7**, 18022

71. Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Ventra, M., Di, Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., *et al* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153