Journal of
Proteomics & Bioinformatics

Wright and Choudhary, J Proteomics Bioinform 2016, 9:6
http://dx.doi.org/10.4172/jpb.1000404

Short Communication Open Access

DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics

James C Wright* and Jyoti S Choudhary

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Abstract

Accurate statistical evaluation of sequence database peptide identifications from tandem mass spectra is essential in mass spectrometry based proteomics experiments. These statistics are dependent on accurately modelling random identifications. The target-decoy approach has risen to become the de facto approach to calculating FDR in proteomic datasets. The main principle of this approach is to search a set of decoy protein sequences that emulate the size and composition of the target protein sequences searched whilst not matching real proteins in the sample. To do this, it is commonplace to reverse or shuffle the proteins and peptides in the target database. However, these approaches have their drawbacks and limitations. A key confounding issue is the peptide redundancy between target and decoy databases leading to inaccurate FDR estimation. This inaccuracy is further amplified at the protein level and when searching large sequence databases such as those used for proteogenomics. Here, we present a unifying hybrid method to quickly and efficiently generate decoy sequences with minimal overlap between target and decoy peptides. We show that applying a reversed decoy approach can produce up to 5% peptide redundancy and many more additional peptides will have the exact same precursor mass as a target peptide. Our hybrid method addresses both these issues by first switching proteolytic cleavage sites with preceding amino acid, reversing the database and then shuffling any redundant sequences. This flexible hybrid method reduces the peptide overlap between target and decoy peptides to about 1% of peptides, making a more robust decoy model suitable for large search spaces. We also demonstrate the anti-conservative effect of redundant peptides on the calculation of q-values in mouse brain tissue data.

Keywords: Shotgun proteomics; FDR; Sequence database; Database searching; Python; Target-decoy

Abbreviations: PSM: Peptide Spectrum Match; FDR: False Discovery Rate; MS: Mass Spectrometry; CID: Collision Induced Dissociation; PPM: Part Per Million; PIT: Percentage Incorrect Targets

Introduction

Shotgun proteomics using tandem mass spectrometry generates accurate peptide molecular weight masses together with fragmentation patterns in the form of spectra for peptides in a biological sample. There are several *in-silico* algorithms to assign spectra to theoretical peptide sequences, predominantly based on searching a protein sequence database, these include Mascot [1], Sequest [2], MS-GF+ [3], and Andromeda [4]. Most software applications report arbitrary fitness scores corresponding to the quality of a peptide to spectrum match (PSM). These scores must then be assessed to allow accurate reporting of experimental accuracy and improve discrimination of true and false identifications. Common statistical metrics reported for proteomic identifications include false discovery rates, q-values and posterior error probabilities (PEP). The false discovery rate (FDR) of an experiment states the estimated percentage of incorrect PSMs at a given significance threshold. Depending of the method used to estimate the FDR it is sometimes possible for a lower significance threshold to produce a better FDR, to address this q-values are used expressing the minimal FDR at which each PSM is significant. Both FDR and q-values represent the amount of error in a dataset, whereas the PEP is the probability of a specific observed PSM identification being incorrect. Calculation of these statistics allows unbiased assessment and comparison of proteomic datasets and the individual PSM identifications. Correctly assessing the false discovery rate of a proteomic dataset is therefore an important step in all experiments. When using sequence database searches to assign peptide sequences to spectra, most approaches will use a set of decoy protein sequences, either concatenated to the search database or as a separate search, with which to estimate the FDR [5-7]. The decoy sequences model the distribution of incorrect peptide sequence matches to query spectra. The assumption can then be made that for every match to a decoy peptide sequence, it is likely there is also a false positive identification of similar score amongst the target PSMs. This assumption requires the decoy sequences to be similar enough to real target protein sequences to represent a random assignment whilst not actually containing any peptides that could be present in the biological sample. However, to accurately model incorrect PSMs a decoy database needs be similar to the target database in terms of size and peptide composition. There are several methods to generate decoy peptide sequences for assessing FDR in the identification of proteomic mass spectrometry data. These include random sequence generation, peptide shuffling, marchov models, and reversing target protein sequences [8]. The most common, easiest and fastest method is to reverse the protein sequences in the target search database. This method preserves peptide cleavage sites and overall composition without making any assumptions about the proteolytic enzyme used to digest the biological protein sample. This method has proven to be a robust and valid method for the majority of small to medium sized experiments. However, for large datasets such as are commonly used in Proteogenomics [9,10], which can include

***Corresponding author:** James C. Wright, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK, Tel: 01223 834244; E-mail: james.wright@sanger.ac.uk

Received May 18, 2016; **Accepted** June 23, 2016; **Published** June 27, 2016

Citation: Wright JC, Choudhary JS (2016) DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. J Proteomics Bioinform 9: 176-180. doi:10.4172/jpb.1000404

many prospective sequences such as translated RNAseq data or full six frame genome translations, there will be a significant number of persistent peptides in the decoy set that are also in the target proteins. These peptides are easily filtered, discarding the overlapping decoy identification from the results prior to estimating FDR or post-processing using software such as Percolator [11,12] or PeptideProphet [13]. However, removing these decoy peptides changes the ratio of decoys to targets and can mask valid decoy assignments. This leads to inaccurate false discovery rate estimation. The effect in smaller scale experiments is marginal; however, the problem is amplified in larger-scale high throughput experiments, as small inaccuracies in the FDR become more significant in larger datasets. This can be further exacerbated when searching large sequence databases or with multiple variable modifications, where the overlap between target and decoy peptides will be greater. This decoy peptide overlap also causes problems beyond PSMs assignment and will be exaggerated at the protein assignment level. One approach to adjust for the peptide overlap is to extend the pi0 or percentage of incorrect targets (PIT) ratio, traditionally used in separated target-decoy searches to adjust FDR estimations to take account for the fact that the majority of targets being true positives. This ratio tries to estimate the true number of incorrect identifications in the target database [6] by modelling the bimodal distribution of target identifications. This correction factor could also be adjusted to account for overlap between the target and decoy databases. However, we propose that reducing redundancy between target and decoy peptides prior to searching is a more effective method and easily achieved by shuffling the overlapping peptides. There will also be a portion of decoy peptides having different amino acid sequences but the same precursor mass and in some cases the same fragmentation pattern as peptides in the target database. This is further compounded when searching with a set of variable modifications. The problem is even greater for high resolution MS data as the accuracy of a precursor mass match is a factor in the scoring of PSMs. To reduce the occurrence of these peptides, cleavage sites can be switched with the preceding amino acid slightly altering the mass of every peptide and reducing the number of matching precursor masses between target and decoy databases. Additionally indistinguishable isobaric amino acids such as leucine and isoleucine need to be considered and can be replaced with a common symbol in both target and decoy sequences [9].

Presented here is a new freely available and open source python tool allowing quick generation of decoy databases for both separate and concatenated searches with a great amount of configurability. DecoyPyrat generates decoys in a hybrid: reverse, switch and shuffle multi-step process as shown in Figure 1. Target protein sequences are reversed and the cleavage sites, defined by the user to suit the experimental data, are switched with the preceding amino acid. At the same time isobaric peptide sequences are resolved replacing leucine and isoleucine with a common amino acid symbol. This process quickly creates an initial reversed and switched decoy database, with

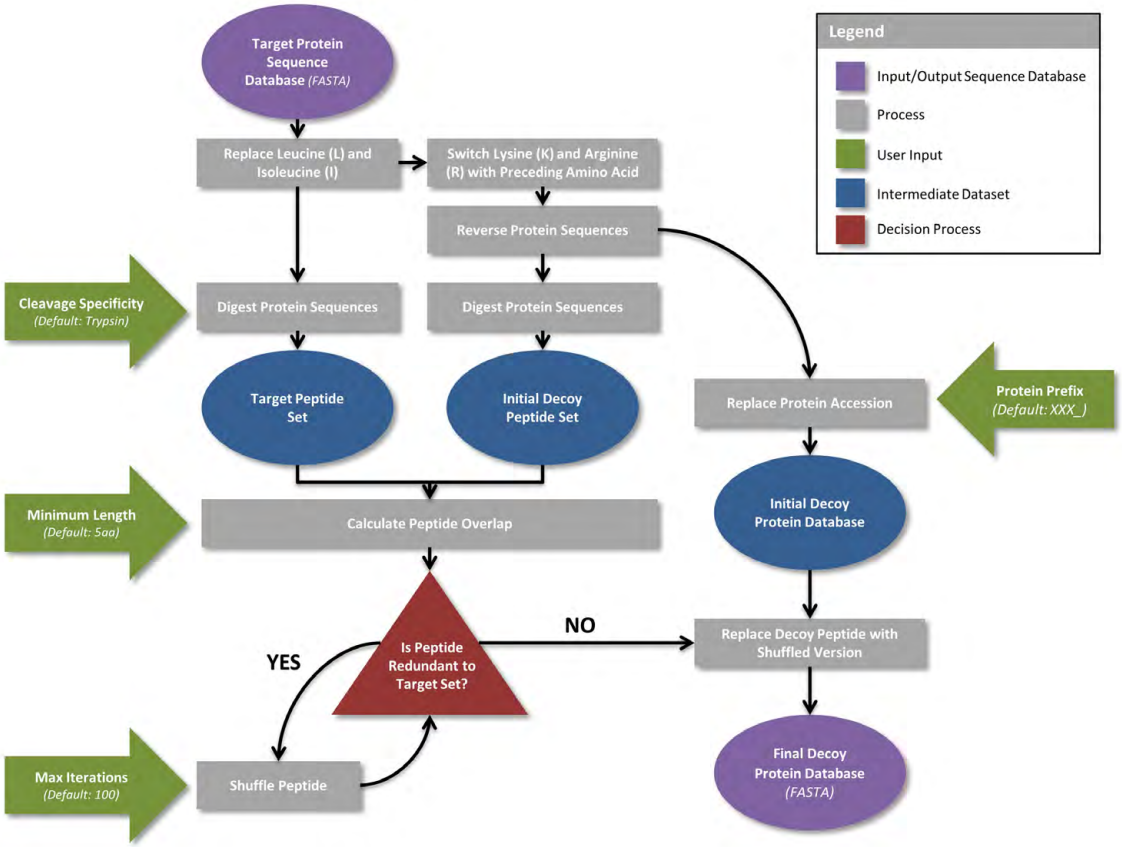

Figure 1: Flowchart for hybrid decoy database generation process in DecoyPyrat. This flowchart describes the process by which DecoyPyrat reads in a FASTA formatted target sequence database replacing isobaric peptides and using a switch and reverse method to generate initial target and decoy digested peptide sets. The overlap between these two sets is examined and overlapping peptides are iteratively shuffled. These peptides then replace their unshuffled counterparts in the final decoy database.

cleavage sites and general peptide composition preserved. The second stage calculates the intersection of peptide sequences between target and decoy peptides. Peptides found to be in both are shuffled iteratively to create a unique decoy peptide. The shuffled peptide sequence then replaces all occurrences of that peptide throughout the decoy database. This preserves the sequence redundancy within the decoy database. To speed up the process a minimum length of peptide is considered when reshuffling, set to 5 amino acids by default as this is the minimum most search engines consider when matching spectra. Occasionally, small low complexity peptides will not have a unique sequence combination, especially with very large target databases. DecoyPyrat reports these failed decoy peptides, which are usually smaller peptides below 7 amino acids in length. To further avoid inaccuracy due to the small number of remaining failed decoy peptides, the minimum peptide length considered and matched to PSMs in a search can be increased; however, this is a trade-off between accuracy and the protein coverage achieved.

Results and Discussion

To assess the performance of DecoyPyrat versus a standard reverse database approach, decoy databases were generated for increasingly larger target databases, obtained from a six frame translation of the mouse reference genome. Some commonly used protein sequence databases including mouse and human reference UniProt proteomes were also included in the analysis. The results of these decoy generation simulations are shown in Figure 2. The runtime for DecoyPyrat performing the hybrid decoy database generation increases in a linear fashion with the size of the target database. It should be noted that the size metric used for the databases is the number of unique peptides, which accounts for any redundancy between the protein sequences with in the target database. The redundant overlap between the target and decoy peptides for the reversed approach increases with the size of the database and can make for more than 5% of the peptides in the larger search spaces. The number of redundant target-decoy peptides is significantly reduced by the hybrid method, maxing out at just over 1% of the total peptides for very large databases, whereas the fraction of redundant peptides in the standard reverse database approach increases to over 5 times this level. This 1% residual redundancy as mentioned previously is mainly small low complexity peptides that do not have a non-redundant amino acid combination. It is also interesting to note that the decoys generated from the complete UniProtKB database, which includes both Swiss-Prot and TrEMBL in their entirety, show much lower peptide redundancy than would be expected for a database of that size based on the trend seen in other databases. A possible explanation for this is the introduction of many more unknown or "X" amino acids in the TrEMBL sequences.

To investigate the effect of the hybrid decoy database method on real search results, a large dataset of publically available CID spectra collected from mouse tissue samples was used. This dataset published in 2013 by Geiger at al., [14] consisted of 136,932 spectra collected on a LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) from mouse brain medulla tissue samples. This dataset was searched using Mascot (v2.5) (Matrix Science) [1] using Mascot's integrated reversed decoy database method. The search parameters used were set to use a full tryptic digest, with up to 3 missed cleavages, 10 ppm precursor tolerance, 0.6 Da fragment tolerance, a fixed carbamidomethyl cysteine modification, and a variable methionine oxidation modification. The search database included all sequences in the M8 GENCODE mouse release [15] concatenated with a full 6 frame translation of the genome filtered to only include open reading frames greater than 20 amino acids. The same dataset was searched again with the same search parameters against separate and concatenated target-decoy databases generated using DecoyPyrat. Using the MascotList results summary utility which is part of the MascotPercolator tool (http://www.sanger.

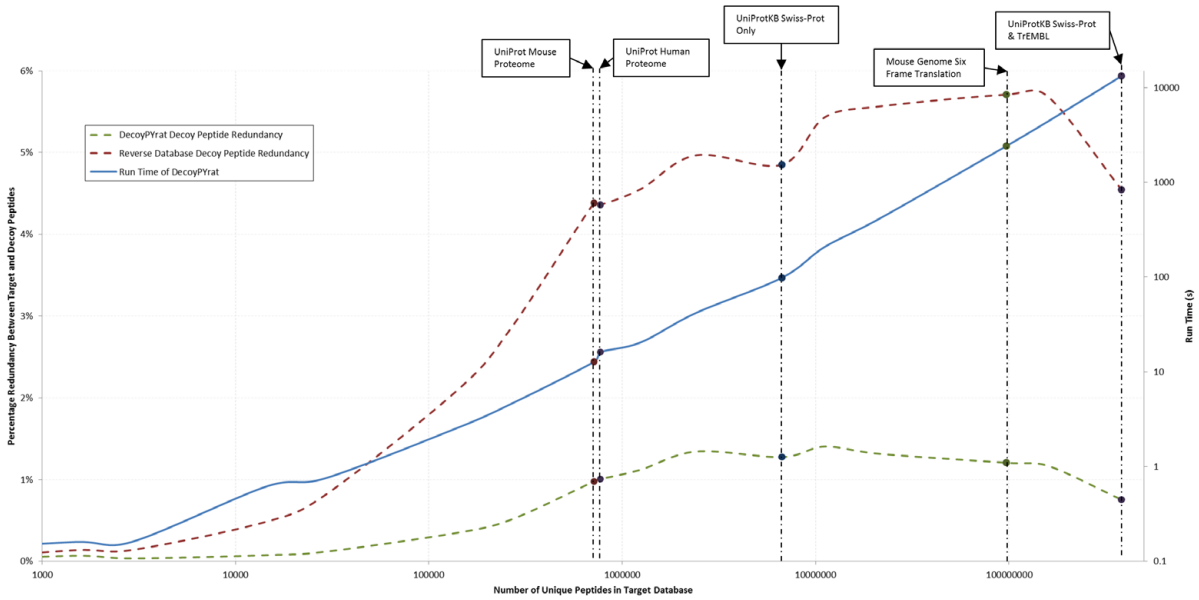

Figure 2: DecoyPyrat runtime and target-decoy peptide overlap. This plot depicts as a percentage the peptide redundancy between the target and decoy databases when using a simple reversing approach versus the hybrid DecoyPyrat method. The x-axis shows the number of unique tryptic peptide sequences in the original target database this roughly scales with the number of proteins in the target database; this axis is displayed on a log10 scale. The primary y-axis displays the percentage of the unique peptide sequences that are common to both the target and decoy databases. The secondary y-axis displays runtime in seconds on a log10 scale. The blue line indicates the runtime of DecoyPyrat.

ac.uk/science/tools/mascotpercolator) [16], target PSM q-values were calculated for each search. These results are displayed in Figure 3. The q-values, equivalent to FDR [17], are directly dependant on the number of decoy and target PSMs at any given score. In the figure the number of PSMs reported at any given q-value threshold are displayed when using a reversed decoy database and separate or concatenated hybrid decoy databases generated with DecoyPyrat. The reverse database method has 10% more significant target PSMs at any given threshold. This may seem like a good thing for reversed decoy databases when trying to obtain as many significant PSMs as possible. However, when considering the fact that the only difference between the searches is the amount of redundancy between the target and decoy databases there is doubt cast on the accuracy of the q-values in the reversed search. The reverse database curve is seen to be shifted to the left, reporting lower q-values for the same number of PSMs, otherwise the methods follow a similar profile.

In conclusion, we present a new tool to quickly generate decoy databases using a hybrid method to reduce peptide redundancy to the target database as well as dealing with isobaric precursor masses. We show this improved method of decoy generation is invaluable for estimating accurate false discovery and error probabilities in large scale proteomics experiments. It overcomes some limitations of the most common approaches to decoy generation that work well for small to medium target search spaces. However, when applied to larger whole proteome or full six frame genome translations as might be used for proteogenomic studies the inaccuracies in the decoy model have a significant effect on the number of PSMs reported at any particular threshold. This has a knock-on effect on the inference of protein level statistics [18]. DecoyPyrat is an efficient open source tool which significantly reduces this inaccuracy in decoy databases; it is fast, adaptable and can be used with many user defined parameters such as digestion specificity. DecoyPyrat is available for download from (http://www.sanger.ac.uk/science/tools/decoypyrat)

Acknowledgements

This work is funded by Wellcome Trust grant (WT098051) at the Wellcome Trust Sanger Institute.

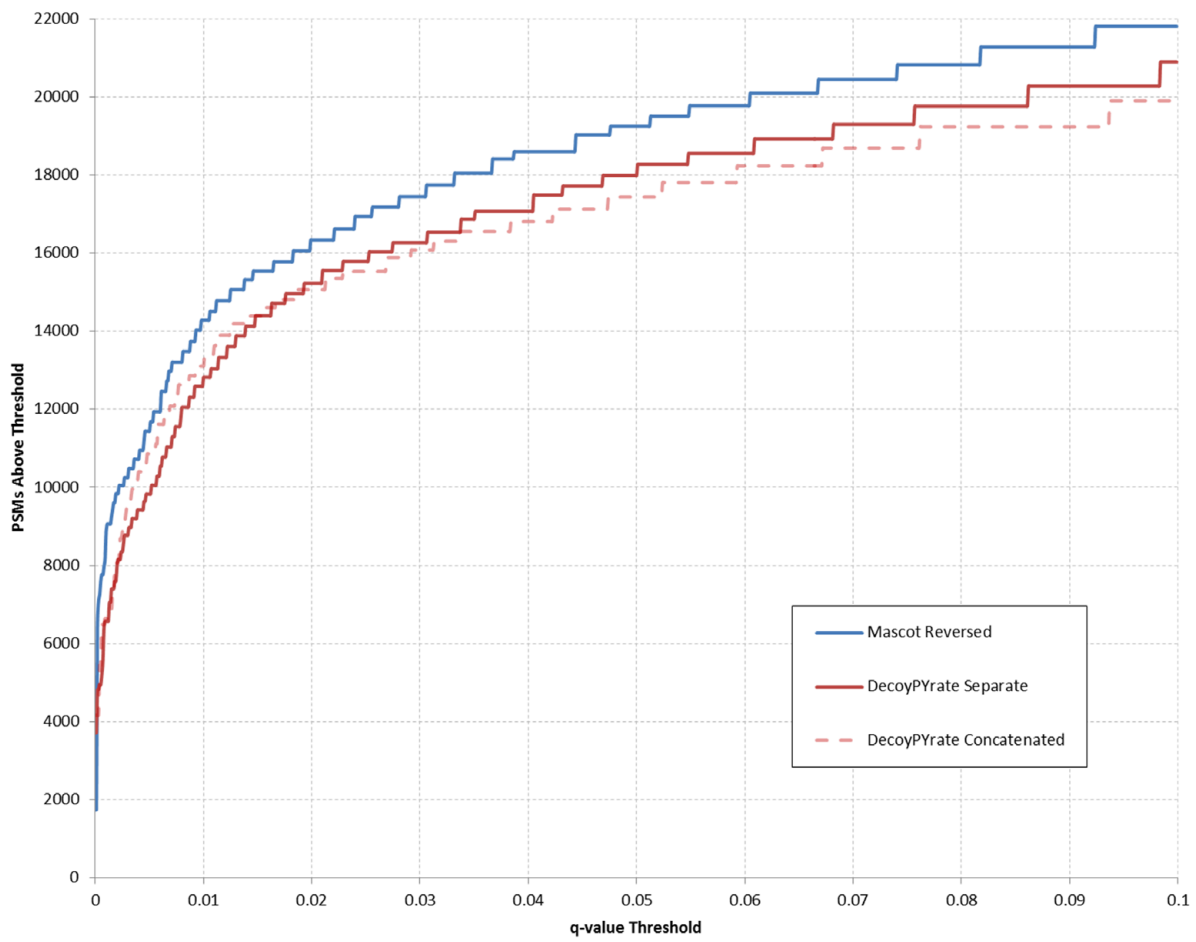

Figure 3: Comparison of significant PSMs reported using reversed database vs. hybrid database. The q-value (FDR) at which PSMs are reported is directly dependant on the number of decoy and target PSMs with scores equal to or above the scoring threshold. In this figure the number of PSMs at any given q-value are reported. The plot compares the curve obtained searching targets with a reverse decoy database (generated using the internal Mascot decoy generation tool) or hybrid decoy databases generated with DecoyPyrat and either searched separately or concatenated with the target database. The reverse database method results in around 10% more PSMs at any given threshold. This may seem good in that more PSMs equates to greater depth of proteome identified. However, when we consider that the only difference between the searches is the redundancy between the target and decoy databases there is doubt cast on the accuracy of the q-value in the reversed search. All three curves follow each other reasonably well with the reverse database shifted to the left and reporting lower q-values.

References

1. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551-3567.
2. Eng JK, McCormack AL, Yates JR (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.
3. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5: 5277.
4. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, et al. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10: 1794-1805.
5. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4: 207-214.
6. Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7: 29-34.
7. Navarro P, Vázquez J (2009) A refined method to calculate false discovery rates for peptide identification using decoy databases. J Proteome Res 8: 1792-1796.
8. Blanco L, Mead JA, Bessant C (2009) Comparison of novel decoy database designs for optimizing protein identification searches using ABRF sPRG2006 standard MS/MS data sets. J Proteome Res 8: 1782-1791.
9. Nesvizhskii AI (2014) Proteogenomics: concepts, applications and computational strategies. Nat Methods 11: 1114-1125.
10. Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, et al. (2016) Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat Commun 7: 11778.
11. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 4: 923-925.
12. Spivak M, Weston J, Bottou L, Käll L, Noble WS (2009) Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. J Proteome Res 8: 3737-3745.
13. Keller A, Nesvizhskii AI, Kolker E, 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.
14. Geiger T, Velic A, Macek B, Lundberg E, Kampf C, et al. (2013) Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. Mol Cell Proteomics 12: 1709-1722.
15. Mudge JM, Harrow J (2015) Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mamm Genome 26: 366-378.
16. Wright JC, Collins MO, Yu L, Käll L, Brosch M, et al. (2012) Enhanced peptide identification by electron transfer dissociation using an improved Mascot Percolator. Mol Cell Proteomics 11: 478-491.
17. Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 7: 40-44.
18. Serang O, Käll L (2015) Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. J Proteome Res 14: 4099-4103.

Citation: Wright JC, Choudhary JS (2016) DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. J Proteomics Bioinform 9: 176-180. doi:10.4172/jpb.1000404