# Tutorial: Using iPtgxDBs (Integrated Proteogenomics search DataBases) to improve the annotation of prokaryotic genomes

Mass spectrometry based proteomics is an analytical method that aims to identify and quantify all proteins present in a biological sample e.g. a tissue, cell line, etc. (Aebersold and Mann, 2003). This field depends upon bioinformatics for data analysis and interpretation (Nesvizhskii et al., 2007). In a standard proteomics workflow, one important step concerns the database search: here, the peaks of the tandem mass spectra (MS/MS) generated by mass spectrometers are matched against theoretical spectra generated from a protein sequence database with the aim to identify the best matching peptide. Fig.1 shows an overview of the basic steps in the analysis of mass spectrometry data. Most commonly, the spectra are searched against a reference protein sequence database, which can be downloaded for example from the NCBI's (National Center for Biotechnology Information) RefSeq (O'Leary et al., 2016) or from Uniprot (The UniProt Consortium, 2017). *De novo* peptide sequencing (Lu and Chen, 2004) is an alternative method that is used when the genome sequence of the organism to be studied is not available (and hence no protein search database).

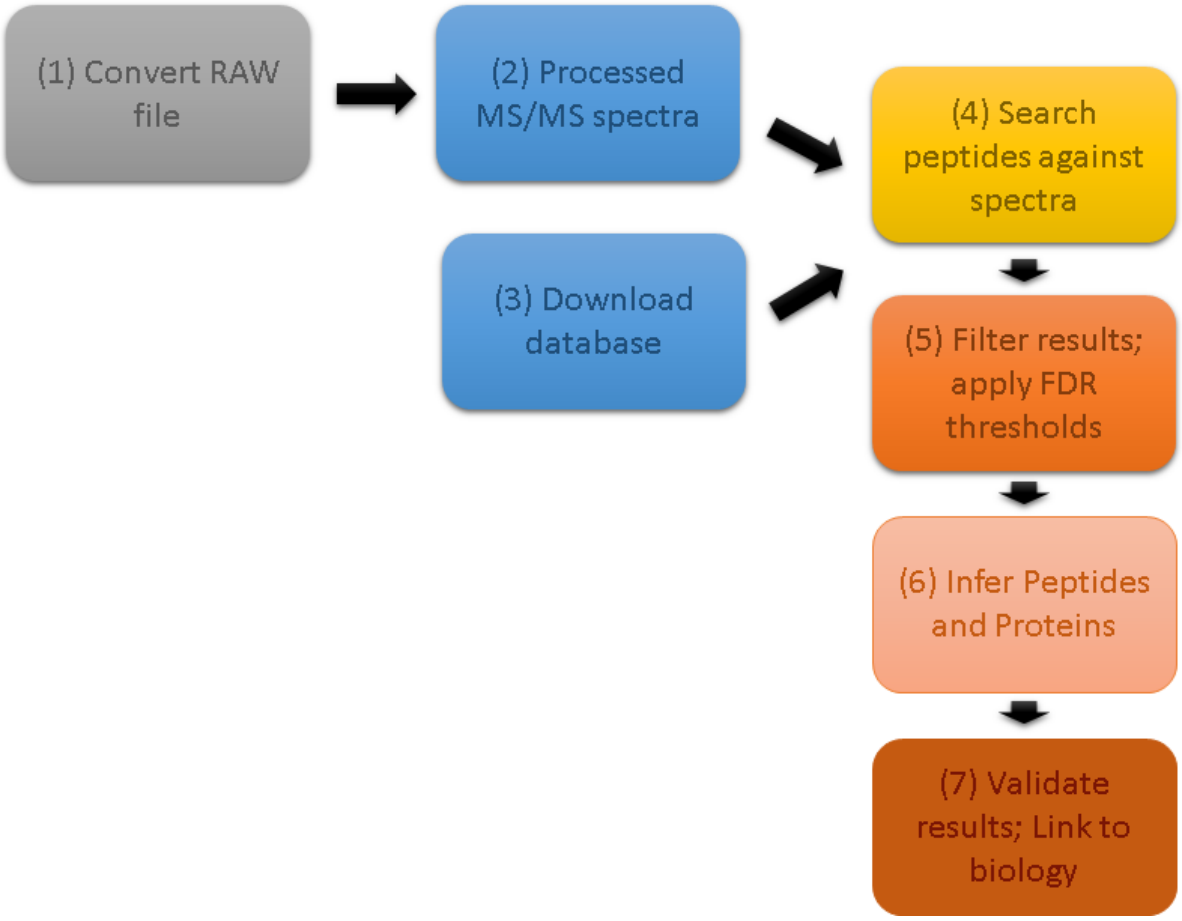


Fig. 1. Basic workflow to analyze proteomics data

**One key problem in proteomics is based on the difficulty to correctly and comprehensively annotate all protein-coding sequences (CDSs) in a sequenced genome**. This, however, is an essential prerequisite to fully exploit the rapidly growing repertoire of completely sequenced prokaryotic genomes. This problem is best illustrated by the fact that different reference annotation resource centers use different approaches and gene prediction algorithms to annotate an identical genome sequence. This leads to large discrepancies among the number of CDSs annotated by different resources, missed functional short open reading frames (sORFs), and over-prediction of spurious ORFs, all of which represent serious limitations (Fig. 2).

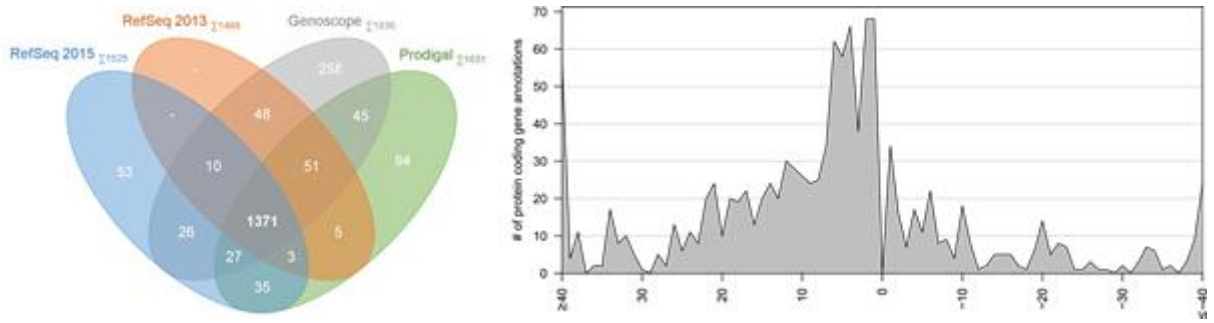


Fig.2: Left panel: A 4 way Venn diagram highlights the differences in the no. of CDSs for *Bartonella henselae* strain houston-1 (a Gram-negative alpha-proteobacterium) predicted by reference genome annotations from RefSeq (including two releases), Genoscope and Prodigal, an *ab initio* gene prediction software. Right panel: Histogram of length differences for all predicted protein start sites from these annotations. Most start site differences affect initiation codons further upstream, leading to longer protein sequences.

Proteogenomics, a research field at the interface of genomics and proteomics, is one attractive approach to address these problems. Here, customized protein sequence databases are generated from genomics and/or transcriptomics sequences, against which proteomics data is then searched (Ahrens et al., 2010; Nesvizhskii, 2014). The left panel of Fig. 3 shows a schematic workflow of a proteogenomic process.

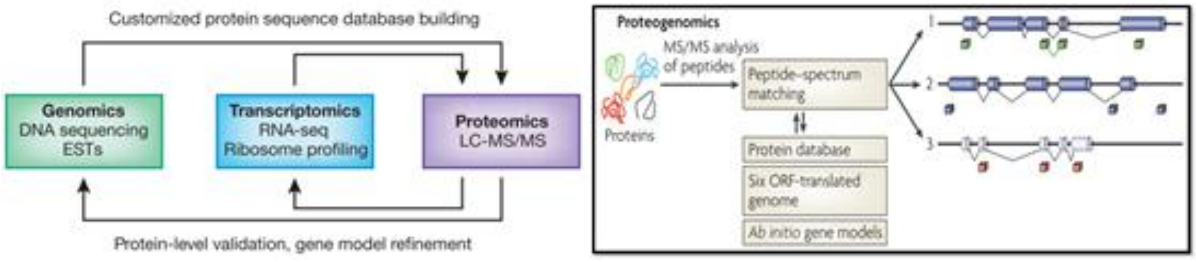


Fig.3: Left panel: Schematic workflow representing a classical proteogenomics process (adapted from (Ahrens et al., 2010; Nesvizhskii, 2014). Right panel: Fragment ion spectra searched against a custom protein database. Peptides identified might map to (1) already known protein coding genes, (2) the close vicinity of known genes, or to (3) not yet annotated regions harboring novel genes that are encoding proteins which may carry out important/essential functions (adapted from (Ahrens et al., 2010; Nesvizhskii, 2014)).

The search against the customized sequence database (DB) can provide direct protein expression evidence for i) already known protein coding genes (CDSs), ii) alternative protein start sites and iii) in some cases even for novel open reading frames (ORFs) that are not annotated in a genome (Fig.2, right panel). **Therefore, proteogenomics can address the key problem of identifying all protein coding genes and their precise start sites described above.** Proteogenomics has been applied to both prokaryotes and eukaryotes ((Ahrens et al., 2010; Nesvizhskii, 2014); (Menschaert and Fenyö, 2015). Several bioinformatics solutions for conducting proteogenomics depend upon a six-frame translation based DB, including Peppy (Risk et al., 2013), Genosuite (Kumar et al., 2013) and PGP (Tovchigrechko et al., 2014). Few other tools integrate data from different species or strains including MScDB (Marx et al., 2013), MSMSpdbb (de Souza et al., 2010), and PG Nexus (Pang et al., 2014). Some solutions alternatively use RNA-seq data to limit the protein search DB size thus achieving better statistical power (Wang et al., 2012; Woo et al., 2014; Zickmann and Renard, 2015). **However, none of the available tools leverages the benefits of manually curated reference annotations and allows to integrate results of ab initio gene predictions and a six-frame translation into one highly informative, non-redundant, and transparent resource.**

Our iPtgxDB (Integrated Proteogenomics DataBase) strategy (Omasits et al., 2017) addresses this unmet need for a robust computation tool and captures the full protein-coding potential of genome sequences. Importantly, it (1) considers results from different reference genome annotations which often include substantial manual curation efforts from experts, and from ab initio gene prediction tools, (2) allows the identification of the small fraction of true functional sORFs often missed by the above annotations or predictions, (3) aid in the annotation of newly sequenced genomes, and (4) enable scientists to visualize their experimental proteomics results in the context of both the genome and all available annotations. Note that **iPtgxDB is designed specifically for prokaryotic genomes.**

In this tutorial, we are going to learn to generate a minimally redundant and maximally informative, integrated proteogenomic database for a genome from up to three different types of annotation sources (Fig. 4) Including reference genome annotation, *ab initio* prediction and *in silico* annotation. We will split the task into three different modules: Module I – iPtgxDB_convert; Module II – iPtgxDB_insilico and Module III – iPtgxDB_combine. Software, example data and results (optional) are provided for each module separately. Please find the files in the respective folders. A separate tutorial help document is included in each of the module folders which describes the steps to be followed. We will use the public web server https://iptgxdb.expasy.org/iptgxdb/submit/ to execute the modules iPtgxDB_insilico and iPtgxDB_combine.

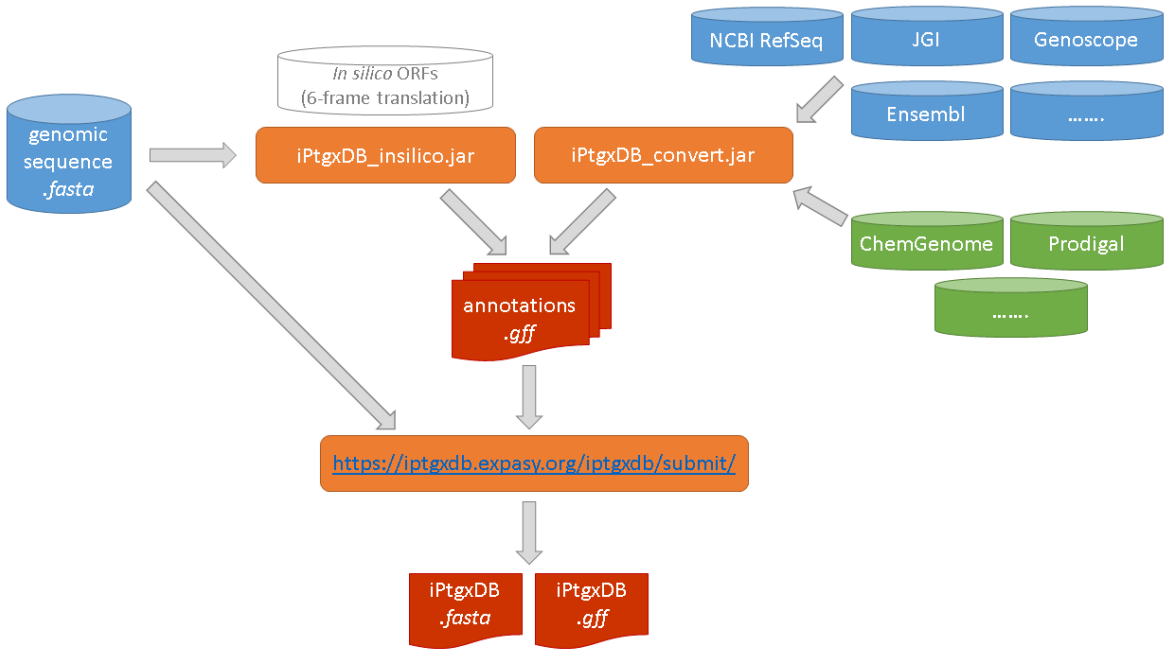


Fig.4: Schematic overview showing the workflow to generate integrated proteogenomics databases using the iPtgxDB tool. The three different modules are shown in the orange boxes. Module I (iPtgxDB_convert) converts the different genome annotations file formats into a generic feature format (GFF) file. Module II (iPtgxDB_insilico) generates a modified six-frame translation of a genome. Module III (iPtgxDB_combine) combines the annotations files from Module I and II into an iPtgxDB search database and a GFF file that provides all integrated annotations, and that can be viewed in a genome browser.

## Module I: iPtgxDB_convert

The software module iPtgxDB_convert converts genome annotation files from popular reference sources (blue containers in Fig.3) and *ab initio* predictions (green containers in Fig.3) into a GFF (https://www.ensembl.org/info/website/upload/gff.html) file, respectively. A GFF file contains information about the annotation features present in a genome including chromosome name, genomic coordinates (start and end), strand, and length of the feature, score and other attributes. GFF files are most commonly used in genomic viewers to visualize all the encoded features in a genomic sequence.

For the prokaryotic genome of interest, reference genome annotations (if available) should be downloaded from NCBI's RefSeq (Pruitt et al., 2012), Genoscope/Microscope (Vallenet et al., 2017), Ensembl (Kersey et al., 2017) and Integrated Microbial Genomes (IMG) initiative of the Joint Genome Institute (JGI) (Markowitz et al., 2014). Any other reference annotation can also be included as far as the genome coordinates for the coding sequences are available. These annotation files leverage the benefit of manual curation on top of *de novo* genome annotation and thus are extremely useful.

On top of the reference genome annotation, *ab initio* genome predictions from Prodigal (Hyatt et al., 2010) and Chemgenome (Singhal et al., 2008) should be obtained. Adding the

ab initio predictions provides benefits for genomes for which reference genome annotation are not available. Specifically, it provides benefit for *de novo* assembled genomes which relies only on ab initio predictions to identify the genes and coding sequences. Prodigal can be run either on the command line or as an installation in Galaxy (https://toolshed.g2.bx.psu.edu/repository). **Note: Most of the genomics based tools are only available for Linux systems and can be used via the command line.**

## Module II: iPtgxDB_insilico

iPtgxDB_insilico is a house created tool that predicts a modified six frame translation after considering alternative start codons. This tool will be highly useful to identify functional sORFs, which are often missed due to rather conservative length thresholds of ab initio predicted ORFs. iPtgxDB_insilico generates the in silico ORF annotation by scanning all six frames of the genome sequence for the longest possible ATG-initiated ORFs. Extensions to these ORFs are considered for the alternative start codons TTG, GTG, and CTG. For regions between two stop codons without ATG codon, the longest alternative start codon initiated ORF is considered. Finally, all potential in silico ORFs above a selectable length threshold (recommended: 18 aa) will be added.

## Module III: iPtgxDB_combine

After all annotations are obtained from Module I and Module II, iPtgxDB_combine will integrate these annotations into a single iPtgxDB which covers the entire protein coding potential of the genome. FASTA and GFF files will be generated for every iPtgxDB. FASTA files can be used with any proteomics search engine and contains highly informative identifier for each sequence that can be easily interpreted and understood. GFF file contains information regarding each sequence in the FASTA file, the annotations that were integrated to generate a sequence, the start and stop position, strand and other attributes which makes the visualization easier in the genome viewer and eliminates the necessity to map back the identified peptides and proteins back to the genome.

A unique aspect of our integrated strategy is that almost all MS-identifiable peptides of the iPtgxDB unambiguously identify one specific protein (Fig. 5). To achieve this, we first extended our PeptideClassifier concept (Qeli and Ahrens, 2010). to prokaryotes: we treat protein sequences with a common stop codon and varying start positions (N termini) as a protein annotation cluster, i.e., an equivalent of a prokaryotic gene model (similar to isoforms of a eukaryotic gene model). Class 1a peptides remain most informative as they are unique to one entry in a DB, while class 1b peptides map uniquely to one annotation cluster with all identical sequences. Class 2a and class 2b peptides map to all sequences of an annotation cluster. Class 3a peptides map to identical sequences from different annotation clusters (typically duplicated genes). Class 3b peptides map to different sequences from different annotation clusters and are least informative. Fig. 5 shows how iPtgxDB_combine integrates the different annotations into an iPtgxDB.

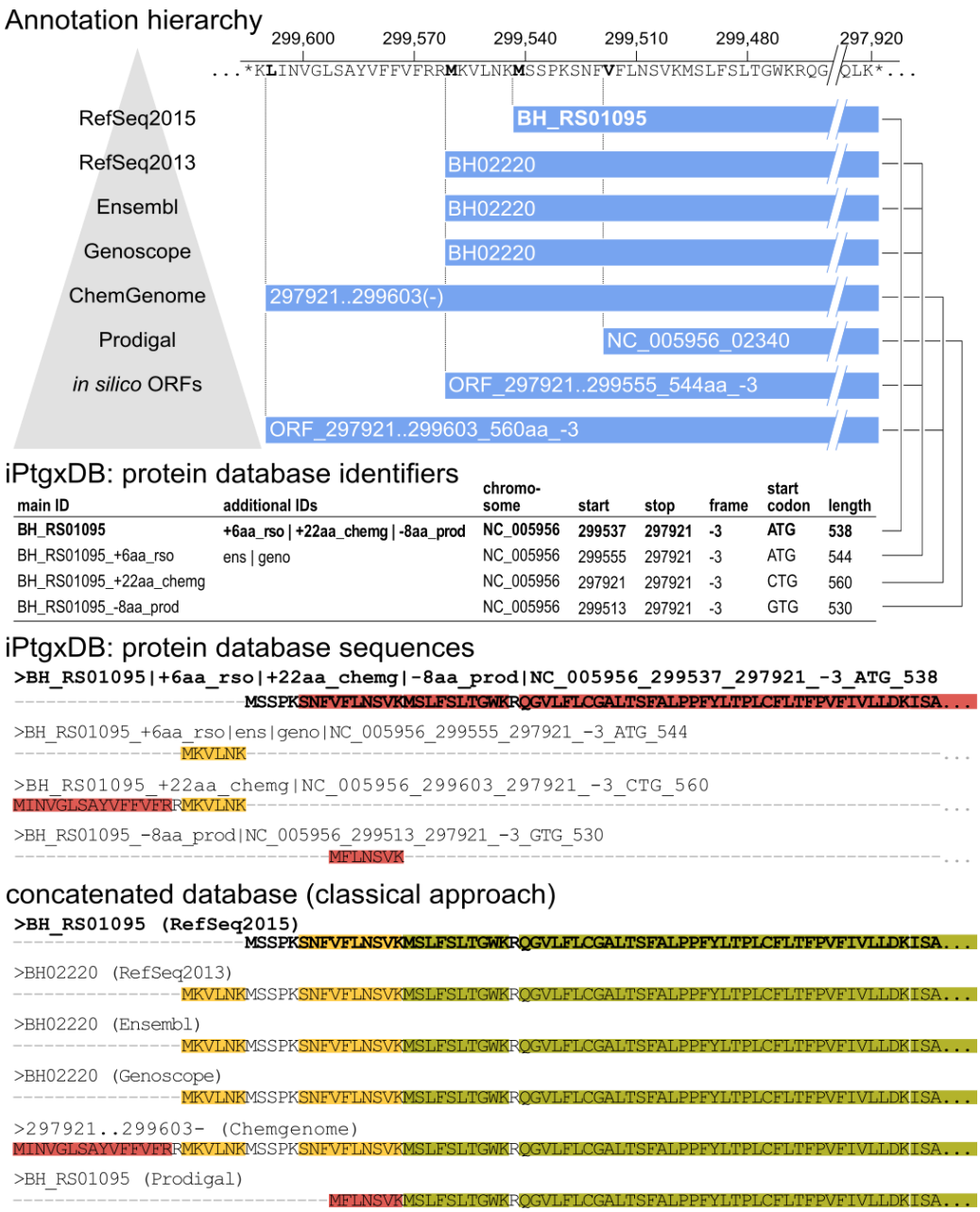


Fig 5: Generating an iPtgxDB with informative identifiers and a minimally redundant protein search DB in FASTA format. CDSs and pseudogenes of seven resources are integrated in a stepwise fashion. Informative protein identifiers are created and illustrated for the annotation cluster, with the RefSeq2015 anchor sequence BH_RS01095 shown in bold, where three additional start sites exist. The four different proteoforms are added to the protein search DB: the anchor sequence (bold) with the full protein sequence, the extensions (RefSeq2013 and ChemGenome) add the upstream sequence up to the first tryptic cleavage site within the anchor sequence. The shorter Prodigal prediction uses an alternative start codon resulting in a distinguishable N-terminal peptide and therefore also gets added. The two in silico ORFs are identical to annotations higher up in the annotation hierarchy and therefore are not added. The colors depict the different peptide classes; red is class 1a peptide which unambiguously identifies one protein/proteoform, orange depicts class 2a peptides that identifies some proteoforms in the annotation cluster but not all and green depicts the class 2b peptide that maps to the identical sequence of all proteoforms and hence is redundant information.

As a first example case study, we will build an iPtgxDB for the prokaryotic model organism *Bartonella henselae* strain Houston-1, called Bhen from here on. Bhen is a Gram-negative $\alpha$-proteobacterium; it is a hemotropic, zoonotic pathogen that can cause cat scratch disease in immuno-competent humans, as well as bacteraemia, endocarditis, and vasoproliferative lesions in immuno-compromised patients (Omasits et al., 2013). In the second example, we will build an iPtgxDB for a *de novo* assembled genome. You will build these iPtgxDB(s) from the annotation files and learn to visualize the GFF file in a genome browser along with the proteomics search results (results will be pre-processed for you) thus realizing the usefulness of iPtgxDB for proteogenomics. To start building the iPtgxDB, go to the folder Module I and follow the instructions given in the PDF document, and move from there to Module II and then finally to Module III.

## Take home message

We present a flexible, yet general iPtgxDB based (proteogenomics) strategy that will allow one to identify novelties beyond reference source based annotations in the genome of prokaryotes of different taxonomic origin (α-, β-, γ-proteobacteria) and widely ranging GC content. Investing a major effort in a preprocessing step to hierarchically integrate reference genome annotations and predictions into an iPtgxDB that covers the entire protein-coding potential pays off: Close to 95% of the peptides unambiguously imply one protein, facilitating swift data analysis and mining. To enable proteogenomics for a larger microbiology research community with access to proteomics core facilities, we provide both a set of precomputed iPtgxDBs for several key prokaryotic model organisms, including founder strains of gene knockout collections, and the software to create them in a public web server (https://iptgxdb.expasy.org/) for any prokaryote (Fig. 6).

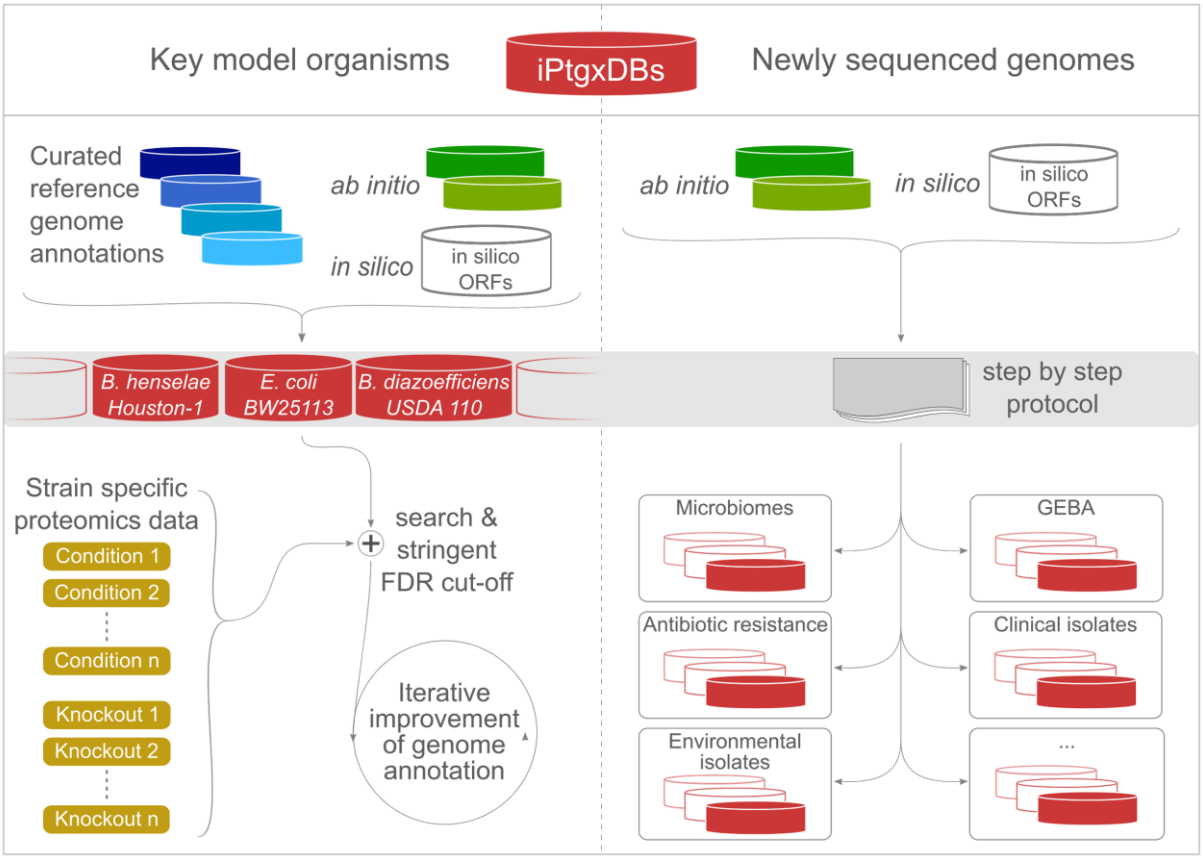


Fig 6. Application of our integrated proteogenomics approach. We release open source iPtgxDBs for several model organisms (https://iptgxdb.expasy.org/); here, for Bhen, E. coli BW25113, and B. diazoefficiens USDA 110 (left panel). Using proteomics data from any condition or knockout strain (light brown boxes; here, schematically shown for E. coli), researchers can identify novelties and iteratively improve the genome annotation, e.g., in a community-driven genome Wiki approach (Salzberg 2007). The release of the software to integrate ab initio predictor(s) and in silico predictions (Supplemental Fig. S8) can help to improve genome annotations of many newly sequenced genomes (right panel).

Enjoy generating your iPtgxDB!

## References


1. Aebersold, R., Mann, M., 2003. Mass spectrometry-based proteomics. Nature 422, 198–207.
2. Ahrens, C.H., Brunner, E., Qeli, E., Basler, K., Aebersold, R., 2010. Generating and navigating proteome maps using mass spectrometry. Nat. Rev. Mol. Cell Biol. 11, 789–801.
3. de Souza, G.A., Arntzen, M.Ø., Wiker, H.G., 2010. MSMSpdbb: providing protein databases of closely related organisms to improve proteomic characterization of prokaryotic microbes. Bioinformatics 26, 698–699.
4. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.
5. Kersey, P.J., Allen, J.E., Allot, A., Barba, M., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Kumar, N., Liu, Z., Maurel, T., Moore, B., McDowall, M.D., Maheswari, U., Naamati, G., Newman, V., Ong, C.K., Paulini, M., Pedro, H., Perry, E., Russell, M., Sparrow, H., Tapanari, E., Taylor, K., Vullo, A., Williams, G., Zadissia, A., Olson, A., Stein, J., Wei, S., Tello-Ruiz, M., Ware, D., Luciani, A., Potter, S., Finn, R.D., Urban, M., Hammond-Kosack, K.E., Bolser, D.M., De Silva, N., Howe, K.L., Langridge, N., Maslen, G., Staines, D.M., Yates, A., 2017. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res.
6. Kumar, D., Yadav, A.K., Kadimi, P.K., Nagaraj, S.H., Grimmond, S.M., Dash, D., 2013. Proteogenomic analysis of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline. Mol. Cell. Proteomics 12, 3388–3397.
7. Lu, B., Chen, T., 2004. Algorithms for de novo peptide sequencing using tandem mass spectrometry. Drug Discov. Today Biosilico 2, 85–90.
8. Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., Anderson, I., Billis, K., Varghese, N., Mavromatis, K., Pati, A., Ivanova, N.N., Kyrpides, N.C., 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res. 42, D560–7.
9. Marx, H., Lemeer, S., Klaeger, S., Rattei, T., Kuster, B., 2013. MScDB: a mass spectrometry-centric protein sequence database for proteomics. J. Proteome Res. 12, 2386–2398.
10. Menschaert, G., Fenyö, D., 2015. Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrom. Rev. 36, 584–599.
11. Nesvizhskii, A.I., 2014. Proteogenomics: concepts, applications and computational strategies. Nat. Methods 11, 1114–1125.
12. Nesvizhskii, A.I., Vitek, O., Aebersold, R., 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods 4, 787–797.
13. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H.,

Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–45.

14. Omasits, U., Quebatte, M., Stekhoven, D.J., Fortes, C., Roschitzki, B., Robinson, M.D., Dehio, C., Ahrens, C.H., 2013. Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. Genome Res. 23, 1916–1927.
15. Omasits, U., Varadarajan, A.R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., Nikolayeva, O., Québatte, M., Patrignani, A., Dehio, C., Frey, J.E., Robinson, M.D., Wollscheid, B., Ahrens, C.H., 2017. An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. Genome Res.
16. Pang, C.N.I., Tay, A.P., Aya, C., Twine, N.A., Harkness, L., Hart-Smith, G., Chia, S.Z., Chen, Z., Deshpande, N.P., Kaakoush, N.O., Mitchell, H.M., Kassem, M., Wilkins, M.R., 2014. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. J. Proteome Res. 13, 84–98.
17. Pruitt, K.D., Tatusova, T., Brown, G.R., Maglott, D.R., 2012. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40, D130–5.
18. Qeli, E., Ahrens, C.H., 2010. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat. Biotechnol. 28, 647–650.
19. Risk, B.A., Spitzer, W.J., Giddings, M.C., 2013. Peppy: proteogenomic search software. J. Proteome Res. 12, 3019–3025.
20. Singhal, P., Jayaram, B., Dixit, S.B., Beveridge, D.L., 2008. Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. Biophys. J. 94, 4173–4183.
21. The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169.
22. Tovchigrechko, A., Venepally, P., Payne, S.H., 2014. PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations. Bioinformatics 30, 1469–1470.
23. Vallenet, D., Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., Mercier, J., Renaux, A., Rollin, J., Rouy, Z., Roche, D., Scarpelli, C., Médigue, C., 2017. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res. 45, D517–D528.
24. Wang, X., Slebos, R.J.C., Wang, D., Halvey, P.J., Tabb, D.L., Liebler, D.C., Zhang, B., 2012. Protein identification using customized protein sequence databases derived from RNA-Seq data. J. Proteome Res. 11, 1009–1017.
25. Woo, S., Cha, S.W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., Bafna, V., 2014. Proteogenomic database construction driven from large scale RNA-seq data. J. Proteome Res. 13, 21–28.
26. Zickmann, F., Renard, B.Y., 2015. MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. Bioinformatics 31, i106–15.