Computational and Structural Biotechnology Journal 20 (2022) 3667–3675

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

Proteotranscriptomics – A facilitator in omics research

Michal Levin *, Falk Butter *

Institute of Molecular Biology (IMB), 55128 Mainz, Germany

ARTICLE INFO

Article history:
Received 15 March 2022
Received in revised form 4 July 2022
Accepted 4 July 2022
Available online 9 July 2022

Keywords:
Gene annotation
Transcriptomics
Proteomics
Proteotranscriptomics

ABSTRACT

Applications in omics research, such as comparative transcriptomics and proteomics, require the knowledge of the species-specific gene sequence and benefit from a comprehensive high-quality annotation of the coding genes to achieve high coverage. While protein-coding genes can in simple cases be detected by scanning the genome for open reading frames, in more complex genomes exonic sequences are separated by introns. Despite advances in sequencing technologies that allow for ever-growing numbers of genomes, the quality of many of the provided genome assemblies do not reach reference quality. These non-contiguous assemblies with gaps and the necessity to predict splice sites limit accurate gene annotation from solely genomic data. In contrast, the transcriptome only contains transcribed gene regions, is devoid of introns and thus provides the optimal basis for the identification of open reading frames. The additional integration of proteomics data to validate predicted protein-coding genes further enriches for accurate gene models. This review outlines the principles of the proteotranscriptomics approach, discusses common challenges and suggests methods for improvement.

Contents

1. Introduction

A gene is basically a sequence of DNA nucleotides that encodes the synthesis of a gene product, which can be either RNA or protein. The annotation of protein coding functional elements in a complex genome is a challenging task that requires not only a highly accurate genome assembly but also the implementation of common gene features such as start and stop codons and splicing signals. The development of highly efficient sequencing technologies enables the sequencing of an ever-growing number of genomes. This necessitates automated gene annotations that are usually highly dependent on the transfer of gene models from related species. While this trend accommodates the need of the progressively omics-oriented science to study gene regulation on a global level in any species of interest, in reality imprecisions in

* Corresponding authors.
E-mail addresses: m.levin@imb.de (M. Levin), f.butter@imb.de (F. Butter).

https://doi.org/10.1016/j.csbj.2022.07.007

annotations can propagate across new assemblies and genes can be missed. In addition, genome assembly and its annotation is still a highly challenging endeavor and thus remains reserved to highly specialized research groups or big consortia. Thus, an alternative approach by which gene predictions are based not on genomic sequences but rather on assembled contigs from RNA-Seq data of polyA enriched mRNA has advantages. As the transcribed part of the genome is devoid of introns and other non-coding sequences the resulting coding-gene predictions are likely more accurate. Adding additional evidence in form of peptide information obtained by mass-spectrometry, a technique broadly used for the identification and characterization of proteins, a Proteo-Transcriptomics Assembly (PTA) workflow can yield high confidence annotation of protein coding genes without the need for genome assembly. RNA-Seq and mass spectrometry data can nowadays be easily produced via on-demand services or just downloaded from RNA-Seq and proteomics repositories and hence the protocols we present and discuss here can even be applied by research groups with no such high-throughput equipment. In the following, we outline the concepts of the approach, discuss common challenges and suggest methods for advancement.

2. Genome annotation remains challenging

Since its development in the late 1970s, DNA sequencing has become one of the most pivotal tools in biomedical research [1]. It initially facilitated the sequencing of whole genomes of phages in the late 1980s [2] followed by several prokaryotic organisms and the first eukaryote, i.e. the yeast *Saccharomyces cerevisiae* in the mid-2000s [3]. Multicellular eukaryotic genome sequencing was achieved soon later starting with the roundworm *Caenorhabditis elegans* [4] and the plant *Arabidopsis thaliana* [5]. As of March 2021, the International Nucleotide Sequence Database Collaboration (INSDC) contained whole-genome DNA sequence information for 6,480 unique eukaryote species, of which only 583 (9%) represented reference-quality chromosome-scale assemblies [6]. The rest remain only draft assemblies, not passing the now generally

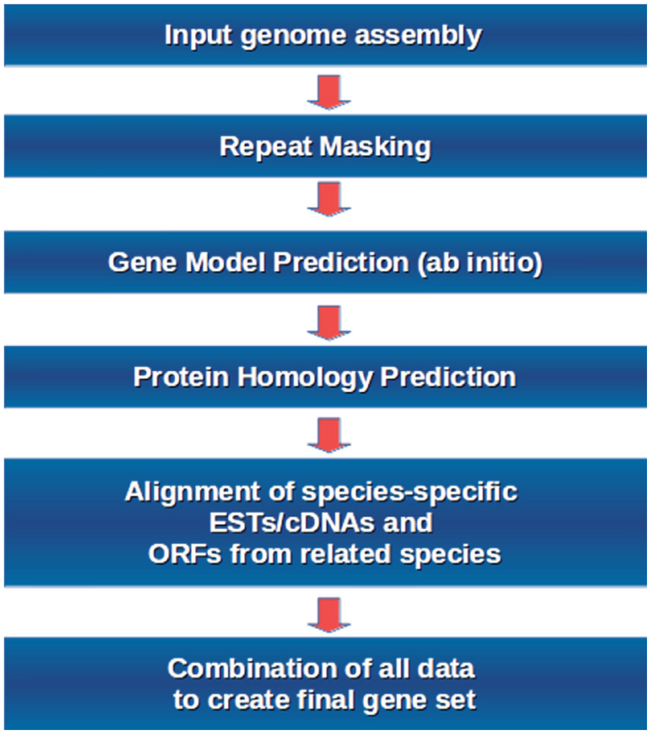
Input genome assembly
Repeat Masking
Gene Model Prediction (ab initio)
Protein Homology Prediction
Alignment of species-specific
ESTs/cDNAs and
ORFs from related species
Combination of all data
to create final gene set

Fig. 2. Main genome annotation steps. Many steps such as repeat masking, protein homology prediction and the alignment of open reading frames from other species include implementing data from other assemblies and annotations and hence mistakes are transferred resulting in impaired precision.

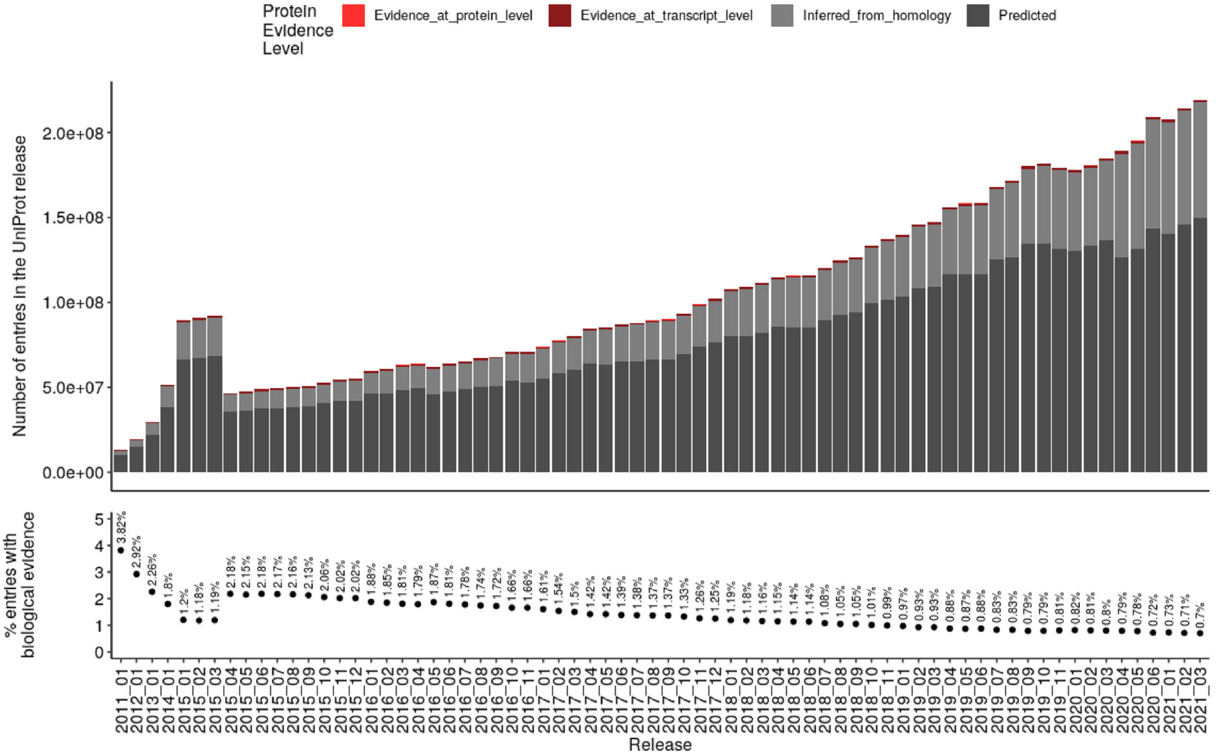

Fig. 1. Paradigm of proteome assembly releases: The number of entries per UniProt Knowledgebase (UniProtKB) [15] release increases extensively with time (upper panel). The vast majority of these entries however (greater than99%) is merely inferred by homology or predicted and has no biological evidence at the transcript or protein level (lower panel). All data presented was extracted from the release notes of the respective UniProt release.

accepted quality requirements [6,7]. Although the number of reference-quality genomes is planned to be increased extensively, this will still require concerted efforts from the community or larger consortia [6,7]. Further, this will not encompass every species of interest and may still require some time.

Accurate and contiguous genomic sequences are important foundations for the identification of functional elements, a process called genome annotation. While this procedure was performed in a highly curated manner with very intense efforts for the first sequenced organisms, the sheer number of sequenced genomes nowadays requires fully automated processes for annotation. In these processes the genome sequence is screened for features of open reading frames that potentially code for proteins. While these measures efficiently enable prediction of possible open reading frames (Fig. 1), the accuracy has been described to suffer in many cases [8,9].

The main challenges of genome annotation can be divided into two categories. The automated annotation of large, fragmented draft genomes still remains very difficult as open reading frames at the edge of contigs as well as in non-assembled genomic regions are lost. This is reflected in the loss of ORFs when compared to fully assembled genomes [10–13]. In addition, draft genome assemblies are known to be frequently contaminated with common bacteria, sequencing vectors, or even human DNA, all of which are ubiquitously present in most labs [14]. These contaminations and any other error in existing annotation i.e. wrongly assigned gene names or a non-genic sequence being annotated as protein coding lead to errors in annotation that tend to propagate across species (Fig. 2). For eukaryotic genomes challenges are even more complex as genes are exceptionally far apart and usually interrupted by introns. That might explain why only 34% of animals with genome assemblies in GenBank also have corresponding annotations [9]. In addition, automated genome annotation mostly provides predictions based on sequence without further evidence unable to control for overprediction (Fig. 1). Hence while genome sequencing technology has continuously improved, genome annotation has become less accurate in general [8].

3. Transcriptome assembly enables gene prediction with reduced complexity

One approach to overcome the obstacles of genome sequence-based gene annotation for protein-coding genes is to start the annotation effort with much less complex underlying data, namely the transcriptome. The transcriptome as the intermediate level of information between genome and proteome is devoid of complex features such as introns and other non-coding sequences and does theoretically constitute the perfect basis for the identification of open reading frames. The large drop in the cost of sequencing also led to the expansion of investigations of transcriptomes of a large range of organisms [16]. This is accomplished by extracting total RNA from the organism of interest, enriching for poly-adenylated transcripts and reverse transcription to create a cDNA library. The cDNA can then be fragmented into various lengths depending on the platform used for sequencing. 454 Sequencing, Illumina, and SOLiD platforms utilize different types of technologies to sequence millions of short reads [17]. Similar to genome assembly the cDNA sequence reads can then be assembled into transcripts. However, established genome assemblers can't be directly used in transcriptome assembly for several reasons. (1) Sequencing depth is assumingly uniform across the genome, while the depth obviously varies between transcripts. (2) In genome sequencing both strands are sequenced, while RNA-Seq is normally strand-specific. (3) Transcript variants of the same locus share different exons and it can be difficult to reconstruct and tease apart all splicing isoforms. Thus, transcriptome assembly has its own challenges. The approach was however strongly enforced by the development of dedicated transcriptome assembly programs [18]. Transcriptome assembly can be performed in two different modes: *de novo* assembly (i.e. assembly of reads without the usage of any reference genome or transcriptome) and genome-guided transcriptome assembly (i.e. reads are mapped to a related reference genome to identify transcript models, which are then assembled into transcripts). While genome-guided assembly provides better results when a well-assembled genome is available, the *de novo* approach enables transcriptome assemblies in cases where genome assembly is absent or in a non-satisfactory shape, i.e. highly gapped or fragmented. Short-read *de novo* transcriptome assemblers generally use one of two basic algorithms: (1) overlap graphs or (2) *de Bruijn* graphs. Overlap graphs are utilized for most assemblers designed for Sanger sequenced reads. The overlap between each pair of reads is computed and compiled into a graph, in which each node represents a single sequence read. This algorithm is more computationally intensive than *de Bruijn* graphs, and most effective in assembling fewer reads with a high degree of overlap. *De Bruijn* graphs align k-mers (sub-sequences within the read with a length of k - usually 25–50 bp) to create contigs. The *de Bruijn* graph approach bypasses the challenge of all-against-all overlap consensus assembly using the full-length reads. While building the graph, the reads are computed as a path through the k-mers and as the k-mers are shorter than the read lengths. This allows fast hashing so the operations in *de Bruijn* graphs are generally less computationally intensive. The following short read assemblers were specially designed for working with RNA-Seq data and are based on *de Bruijn* graphs: Trans-ABySS [19], Trinity [20,21], Oases [22], IDBA-Tran [23], SOAPdenovo-Trans [24], and Shannon [25]. Bridger [26] and BinPacker [27] are two assembly tools that rely on splicing graphs [26] instead of *de Bruijn* graphs. SPAdes v3.13.0 [28] is a widely used *de novo* genome assembler based on *de Bruijn* graphs and MK values. These assemblers have been used to provide transcriptomes for chickpea [29], planarians [30], Parhyale hawaiensis [31], as well as the Nile crocodile, the corn snake, the bearded dragon, and the red-eared slider [32], to name just a few.

Although the *de novo* mode facilitates the inference of many valid and precise transcripts, the approach also bears some potential issues, namely: possible assembly errors in paralogs and multigene families; production of errorsome chimeras; problems reaching full transcript length, and misestimation of allelic diversity [33–35]. Using short read sequences for transcriptome assembly sometimes suffers from low accuracy, especially for the transcripts from eukaryotes that contain complex isoforms [35,36]. This can be partially tackled by using long-read sequences which span longer parts of the original transcript and hence allows for more precise assembly [37]. A downside to long-read sequencing is that the accuracy per read can be much lower than that of short-read sequencing introducing other errors into assembled contigs. Lately hybrid approaches integrating both short and long read sequences in the assembly process have been proposed [38]. To benchmark the transcriptome assemblies overall and the assembled contigs individually, control measures and programs that can characterize these have been developed. Mapping rate (re-mapping rate can give preliminary insights into the quality of a transcriptome assembly), Ex90N50 statistics (expression-informed ExN50 statistic), rate of full-length protein-coding transcripts reconstruction, rnaQUAST [39] (completeness and correctness levels of the assembled transcripts), TransRate [40] (confidence and completeness measures based on the reads used for the assembly only), DETONATE [41] (compactness of the assembly and its support from the RNA-Seq reads) and BUSCO [42] (abundance of single-copy orthologs in the assembly) are some of the more established measures and tools. Using these tools the general quality of the assembly can be measured, and some of the algorithms (TransRate, DETONATE) also provide per-transcript measures that allow further filtering of assembled transcripts to keep only high confidence transcripts.

3.1. Integration of peptide evidence increases confidence in protein predictions

Although transcriptome assemblers get quicker and work with increasing precision, transcript isoform variation contributes to transcriptome complexity and ultimately the quality of transcriptome assemblies. The consequence is overprediction and misassembly especially in loci with high levels of alternative splice forms, allelic variants, close paralogs, close homologs, and close homeologs. Using assembly quality assessment tools as mentioned above, some of these misassembled transcripts can be identified and filtered out. However incorrect frame-shifted open reading frames can only be detected by either comparison to known well evidenced proteins from other species or by using evidence at the protein sequence level. This can be accomplished by cross checking the predicted open reading frame pool from transcriptome assembly with mass-spectrometry peptide identifications. Using this approach, open reading frames stemming from misassembled transcripts can be eliminated by establishing evidence at the protein level and herewith strengthening the confidence in the predictions. Similar approaches have been used to identify non-canonical proteins and novel alternative splicing isoforms which would be lost when working with predetermined annotation databases [43–45]. The principle is easy; RNA-Seq data is assembled to possible transcripts, these are the basis for the prediction of potential open reading frames, which are then used as search space for mass spectrometry peptide information (Fig. 3). This Proteo-Transcriptomics Assembly (PTA) approach enables unbiased proteome annotation without the need for genome information. The ultimate result of the PTA process are transcript contigs bearing open reading frames that are validated by the presence of peptides and hence represent a set of high confidence protein coding transcripts. Preferably the same sample may be used for preparing RNA and protein extracts, however any available raw data (RNA-Seq or mass spectrometry data) can also be downloaded from official repositories like GEO [46], SRA [47] or PRIDE [48] and be combined retrospectively. This opens avenues for the annotation of high confidence open reading frames facilitating research in a cost-effective approach to improve previous or generate new gene models. As the technique can be based on *de*

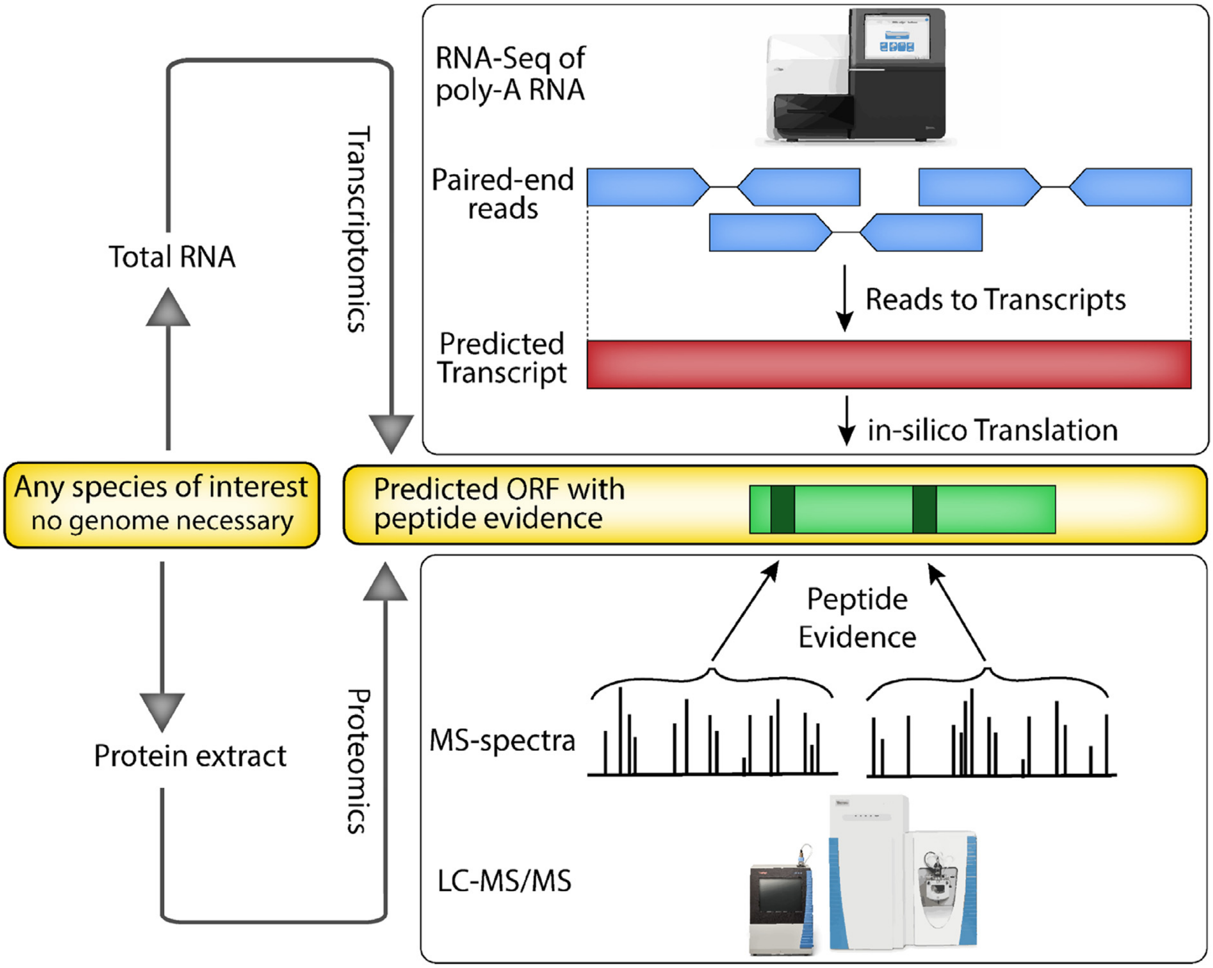

Fig. 3. General outline of the PTA (Proteo-Transcriptomics Assembly) approach. RNA-sequencing data of all poly-adenylated RNA molecules of any species of any cell origin is used for transcript assembly in which individual reads are concatenated into potential full-length transcript contigs. The predicted transcript contigs are then *in-silico* translated into predicted protein sequences in all possible frames. These predictions are used to find potential open reading frames taking important features of common protein coding transcripts (such as a Methionine start and an in-frame stop codon) into consideration. In parallel the proteome of the same sample used for RNA-sequencing is measured with a high-resolution mass spectrometer. The mass spectrometer first records the mass/charge (*m*/*z*) of each peptide ion and then selects the peptide ions individually to obtain sequence information via MS/MS. Peptide fragmentation spectra are matched to in silico generated peptide fragmentation patterns. The ultimate result of the process are transcript contigs that were validated by the presence of peptides and hence represent a set of high confidence protein coding transcripts.

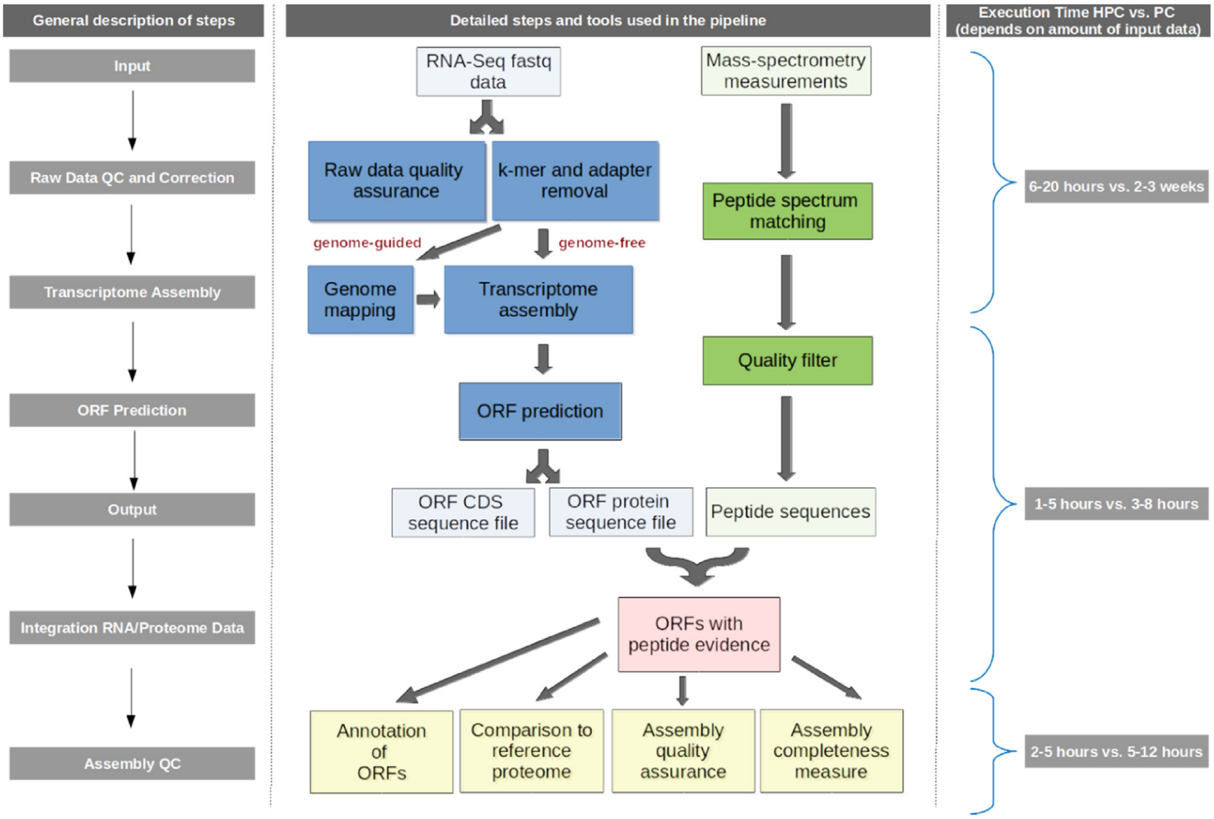
General description of steps
Input
Raw Data QC and Correction
Transcriptome Assembly
ORF Prediction
Output
Integration RNA/Proteome Data
Assembly QC
Detailed steps and tools used in the pipeline
RNA-Seq fastq data
Mass-spectrometry measurements
Raw data quality assurance
k-mer and adapter removal
Peptide spectrum matching
genome-guided
genome-free
Genome mapping
Transcriptome assembly
Quality filter
ORF prediction
ORF CDS sequence file
ORF protein sequence file
Peptide sequences
ORFs with peptide evidence
Annotation of ORFs
Comparison to reference proteome
Assembly quality assurance
Assembly completeness measure
Execution Time HPC vs. PC (depends on amount of input data)
6-20 hours vs. 2-3 weeks
1-5 hours vs. 3-8 hours
2-5 hours vs. 5-12 hours

Fig. 4. General outline of the PTA workflow. In blue: RNA-Seq data preparation steps include 1. the validation of sufficient quality of the sequencing data (**FastQC** [52], fastqp [53], fastq-stats [54]); 2. raw RNA-Seq reads correction and adapter removal (**Rcorrector** [55], QuorUM [56], specialized scripts from **TranscriptomeAssemblyTools** (FilterUncorrectablePEfastq.py); **TrimGalore** (a wrapper around Cutadapt [57] and FastQC [52]); 3. Mapping of reads to a reference genome for the genome-guided mode (**STAR** [58], Bowtie2 [59], BWA [60], Hisat2 [61], TopHat2 [62]); 4. Transcriptome assembly (**Trinity** [20,21], Oases [22], Trans-ABySS [19], SOAPdenovo-Trans [24], IDBA-Tran [23], Bridger [26], BinPacker [27], Shannon [25], SPAdes-sc [28], SPAdes-rna [28]); 5. Identification of candidate coding regions within reconstructed transcript sequences from the previous step (**TransDecoder** [21], FrameD [63], GeneMarkS [64]). In green: mass spectrometry spectra processing and filtering (**MaxQuant** [65], ProteomeDiscoverer (Thermo Scientific), FragPipe [66], MS-GF+ [67]). In red: The predicted ORF protein sequences will be used as search space for the identified peptides extracted from MS/MS spectra. In yellow: ORFs with peptide evidence can be functionally annotated (**Trinotate** [68], blast2GO [69], annot8r [70], Annoscript2 [71]). Newly established annotations can be compared with current annotations e.g., from UniProt and Ensembl (**blastp** [72], DIAMOND [73]), checked for assembly quality standards (**TransRate** [40], rnaQUAST [39], Detonate [41]) and examined for proteome completeness (**BUSCO** [42]). Programs that can be used for the individual steps are listed, while the ones that were tested to work well and deliver satisfactory results in our hands are bolded. The list, though being comprehensive, is not intended to be complete. Beyond the tools listed, alternative tools that may work equally well may exist or being developed. The right panel depicts the computation times of the different steps compared between High-Performance-Computing machines and strong tabletop PCs. The times are only representative, based on the tools marked bold, and depend on the amount of raw data processed and the underlying computing architecture. Execution time may vary for alternative tools used for the individual steps.

novo transcriptome assembly, it provides the possibility to study any species also in the absence of genome sequence information enabling gene discovery, comparative analyses, estimation of expression abundances, and identification of sequence variants.

4. Proteo-Transcriptomics assembly – Challenges

4.1. Computational complexity

While PTA delivers promising results, the implementation of the various programs, tools and custom scripts is not a straightforward endeavor yet. A possible workflow (outlined in Fig. 4) for the full analysis including QC requires the implementation of at least 10 different programs. As there is no pipeline available yet, the application of the approach remains reserved to computationally experienced researchers, limiting it to more highly relevant fields. One possible solution to this issue would be the implementation of a workflow framework that eases the writing of data-intensive computational pipelines, e.g. Nextflow [49], Snakemake [50] or bpipe [51] to automate all relevant programming steps in a parallelized preferably portable pipeline. This would enable a scalable and reproducible analysis also for research groups with less computational experience.

An additional bottleneck for the more widespread application of the PTA approach is the relatively high demand of computing resources of the individual processes, especially of the transcriptome assembly step. As a consequence, the process cannot be efficiently performed on every computational platform (Fig. 4) and is still beyond regular desktop PCs. One solution to this problem is the execution of the workflow on a server or within a High-Performance Computing (HPC) infrastructure. Today, many academic and scientific research firms that require massive computational processing power also use cloud computing instead of establishing their own computing infrastructure. HPC via the cloud can be expanded, adapted and shrunk on demand with affordable costs. For the overall workflow, the implementation of HPC specific features facilitates most efficient execution in terms of processing time and resources.

4.2. Transcriptome assembly accuracy

A known issue with all transcriptome assembly programs is a more or less severe level of fragmented contig assembly. Such frag-

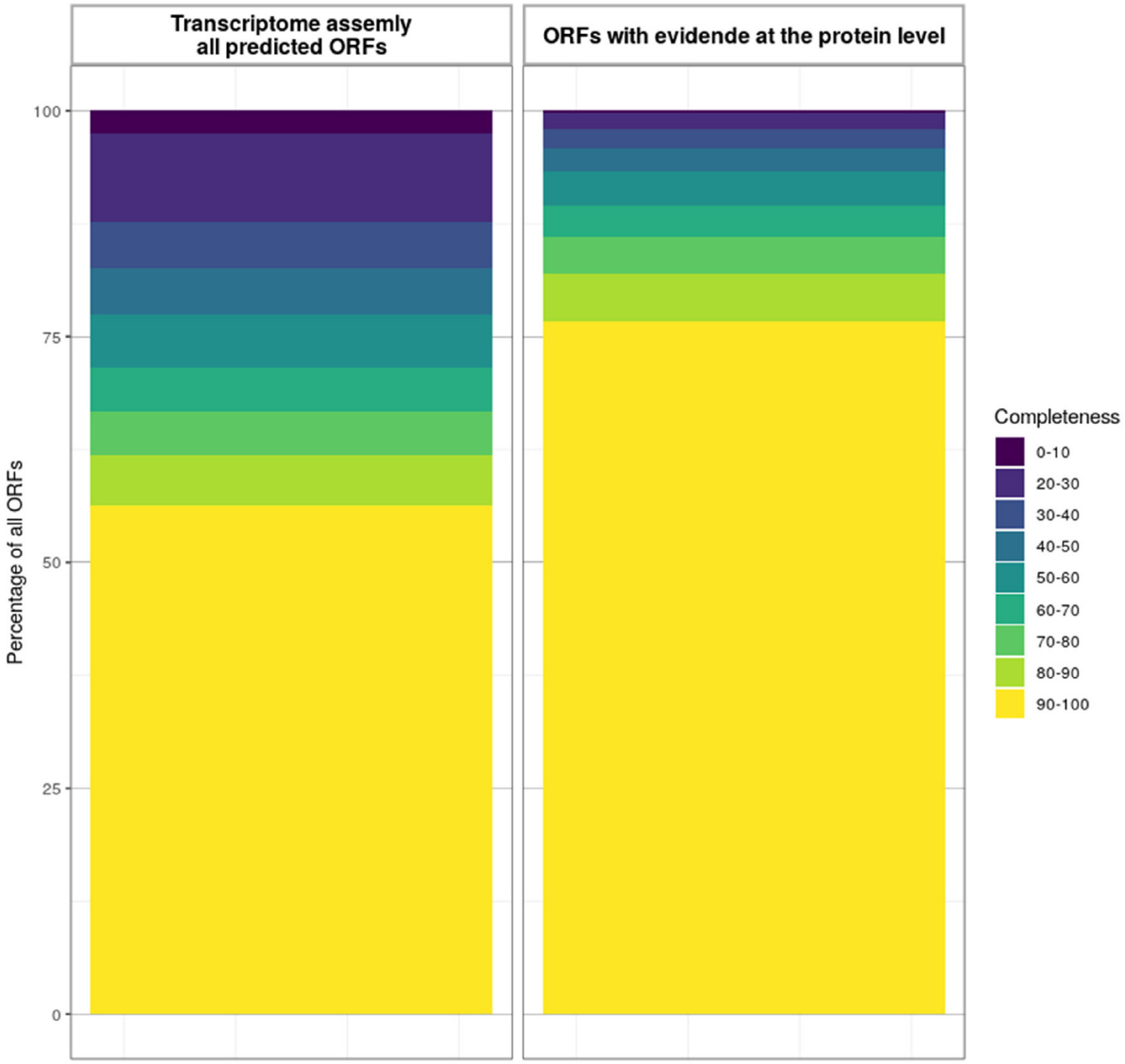

Fig. 5. Open reading frames can be predicted from the assembled transcripts. A known issue of transcriptome assembly is that under certain circumstances (see details in main text) the assembler is not able to assemble the complete transcript but the assembled transcript rather represents a fragment of the actual transcript. The completeness can be measured by comparing the assembled transcripts to current annotations. Depicted are the proportions of assembled transcripts in our previously published transcriptome assembly of the silkworm *Bombyx mori* [74] with different levels of completeness when compared to the genome-based annotation of the silkworm from SilkBase [75]. The left panel represents the distributions in all raw transcript assemblies. Only around 62% of the transcripts show completeness of more than 80%. However, in the pool of predicted open reading frames that could be verified at the protein level (depicted in the right panel) the proportion of near complete transcripts increases to 82%. These gene annotations with additional peptide evidence are enriched for full-length transcripts and thereby increase accuracy.

mented contigs lack a start or a stop codon, or both and hence represent only partial open reading frames and lead to noisy results. The main reasons for partially assembled contigs are low read coverage at a locus, repetitive regions, differential expression of different exons, polymorphism, and sequencing errors, which might potentially lead to local assembly errors. The most efficient way to clean assemblies from these false contigs would be to use measures that would detect any of the underlying causes and then try to filter contigs with high chance of being a wrong assembly, keeping only high-confidence full-length contigs. There are two programs that facilitate the detection of such features. Both TransRate [40] and Detonate [41] provide metrics which take the mapping of reads against the contigs into account in assessing the assembly quality. In addition to an overall assembly score for a given assembly, for each contig within the assembly, TransRate [40] and Detonate [41] provide a score that assesses how well that contig is supported by the RNA-Seq data and that can be used to filter suspicious contigs. While using these measures can help to enrich for high confidence predictions, we observed that the pool of predicted proteins for which peptide evidence can be detected, the overall completeness of the assembled transcripts seems to be significantly higher (Fig. 5). This also emphasizes the importance of adding peptide evidence to predictions, a step most current genome annotations lack (Fig. 1).

4.3. Considerations of proteome coverage

Addition of protein data will increase confidence for the existence of an assembled transcript. Most proteomic data is available as peptide identifications from bottom-up experiments and can be accessed on databases like PRIDE [48] and Massive [76]. However, while high confidence peptide identification has been aided by ever more accurate mass spectrometers in the last decade, currently even for in-depth proteomes, we unfortunately only measure peptides of the more abundant proteins in a sample [77]. This naturally limits the PTA approach as only a fraction of the predicted open reading frames can thus be supported by peptide evidence. Despite this current limit, increases in proteome coverage can further enhance the comprehensiveness of PTA. Advances in mass-spectrometry instrumentation [78–80] and acquisition methods [81–83] enable increasing measurement depth. The use of specific methodology like removal of high-abundant proteins or fractionation approaches can split sample complexity across the measurement and are readily implementable [84]. In addition, the use of samples from different developmental stages, tissues or treatments can modulate and increase the pool of expressed proteins allowing to obtain more peptide evidence [85].

5. Summary and outlook

Identifying all coding regions in a genome is crucial for any study at the level of molecular biology, ranging from single-gene cloning to genome-wide measurements using RNA-Seq or mass spectrometry. While satisfactory annotation has been made feasible for well-studied model organisms through great efforts of big consortia, for many species this kind of data is either absent or not adequately precise. We here reviewed an approach that seeks to overcome many of the bottlenecks of detecting protein-coding regions in the genome. We could previously show that by combining in-depth transcriptome sequencing and high resolution mass spectrometry by proteotranscriptomics we achieved improved gene annotation of protein-coding genes in the *Bombyx mori* cell line BmN4, which is an increasingly used tool for the analysis of piRNA biogenesis and function [74]. Using the PTA approach, we provided the exact coding sequence and evidence for more than six thousand expressed genes on the protein level. This approach outperformed current *Bombyx mori* gene annotation efforts from 4 different sources in terms of accuracy and coverage [74]. Similar approaches were also successfully applied by other groups in various different species and fields such as in human placental samples [86] and leukemia cells [87] and for the detection of microproteins in human [88], in rat [89], pigs [90], mosquitos [91,92], in a combined analysis of human and adenovirus [93], and plants such as the opium poppy [94] and *Michelia maudiae* [95] demonstrating that proteotranscriptomics is widely applicable.

The presented PTA approach can in principle be applied by any individual lab and without prior genomic information. Although most labs do not have their own next-generation sequencer or a high-resolution mass-spectrometer, access to these services from different in-house or external providers are easily available. In principle, even already existing data from different RNA-Seq and proteome repositories can be incorporated eagerly well for PTA.

As mentioned above a significant bottleneck of PTA is the computational complexity of the different bioinformatic analysis steps, which also need considerably large computing resources. These obstacles can be overcome by building a computational pipeline that executes the different processes in a highly parallelized and streamlined manner on an HPC platform or in the cloud. Indeed, we are currently developing a workflow that will be deployable in cloud computing infrastructure and will make benchmarked PTA feasible for anyone interested. Another common issue, fragmented transcript assemblies, has been the source for the development of quality control programs that provide quality measures for assembled contigs. In the future, we envision integrating this QC information with a machine learning algorithm to facilitate identifying potentially fragmented transcript assemblies even more precisely.

In summary, Proteotranscriptomics is an efficient, cost-effective and accurate approach to improve previous gene annotations or generate completely new gene models. As this technique is based on *de novo* transcriptome assembly, it provides the possibility to study any species also in the absence of genome sequence information, for which proteogenomics in its stricter meaning is impossible. Easier computational access and solving major bottlenecks such as program application, efficient transcriptome assembly and automatic quality controls are the next steps to make this approach feasible and reproducible for the broader scientific community.

CRediT authorship contribution statement

Michal Levin: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. **Falk Butter:** Conceptualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Falk Butter reports administrative support was provided by German Research Foundation.

Acknowledgements

We gratefully acknowledge the GenEvo RTG funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2526/1 – Projectnr. 407023052 enabling the conception of this review.

References

[1] Roberts J, Middleton A. Genetics in the 21st century: Implications for patients, consumers and citizens. F1000Research 2018;6. https://doi.org/10.12688/F1000RESEARCH.12850.2/DOI.

[2] Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 1976;260:500–7. https://doi.org/10.1038/260500A0.

[3] Goffeau A, Barrell G, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science 1996;274:546–67. https://doi.org/10.1126/science.274.5287.546.

[4] Consortium C, Elegans S. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 1998;282:2012–8.

[5] Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, et al. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nat 2000;2000 (4086814):796–815. https://doi.org/10.1038/35048692.

[6] Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A 2018;115:4325–33. https://doi.org/10.1073/PNAS.1720115115.

[7] Lewin HA, Richards S, Aiden EL, Allende ML, Archibald JM, Bálint M, et al. The Earth BioGenome Project 2020: Starting the clock. Proc Natl Acad Sci U S A 2022;119. https://doi.org/10.1073/PNAS.2115635118/SUPPL_FILE/PNAS.2115635118.SAPP02.PDF.

[8] Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biol 2019;20:1–3. https://doi.org/10.1186/S13059-019-1715-2/METRICS.

[9] Hotaling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: Where are we now? Proc Natl Acad Sci U S A 2021;118. https://doi.org/10.1073/PNAS.2109019118.

[10] Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. PLoS One 2011;6. https://doi.org/10.1371/JOURNAL.PONE.0021400.

[11] Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 2021;592:737–46. https://doi.org/10.1038/S41586-021-03451-0.

[12] Manchanda N, Portwood JL, Woodhouse MR, Seetharam AS, Lawrence-Dill CJ, Andorf CM, et al. GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics 2020;21. https://doi.org/10.1186/S12864-020-6568-2.

[13] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210–2. https://doi.org/10.1093/BIOINFORMATICS/BTV351.

[14] Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol 2020;21. https://doi.org/10.1186/S13059-020-02023-1.

[15] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49:D480–9. https://doi.org/10.1093/NAR/GKAA1100.

[16] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57. https://doi.org/10.1038/NRG2484.

[17] Slatko BE, Gardner AF, Ausubel FM. Overview of Next Generation Sequencing Technologies. Curr Protoc Mol Biol 2018;122:e59.

[18] Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience 2019;8:1–16. https://doi.org/10.1093/GIGASCIENCE/GIZ039.

[19] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods 2010;2010(711):909–12. https://doi.org/10.1038/nmeth.1517.

[20] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29:644–52. https://doi.org/10.1038/nbt.1883.

[21] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;8:1494–512. https://doi.org/10.1038/nprot.2013.084.

[22] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 2012;28:1086–92. https://doi.org/10.1093/BIOINFORMATICS/BTS094.

[23] Peng Y, Leung HCM, Yiu SM, Lv MJ, Zhu XG, Chin FYL. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics 2013;29:i326–34. https://doi.org/10.1093/BIOINFORMATICS/BTT219.

[24] Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 2014;30:1660–6. https://doi.org/10.1093/BIOINFORMATICS/BTU077.

[25] Kannan S, Hui J, Mazooji K, Pachter L, Shannon TD. An Information-Optimal de Novo RNA-Seq Assembler. BioRxiv 2016;39230. https://doi.org/10.1101/039230.

[26] Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: A new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol 2015;16:1–10. https://doi.org/10.1186/S13059-015-0596-2/FIGURES/7.

[27] Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, et al. BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. PLOS Comput Biol 2016;12:e1004772.

[28] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012;19:455. https://doi.org/10.1089/CMB.2012.0021.

[29] Kudapa H, Azam S, Sharpe AG, Taran B, Li R, Deonovic B, et al. Comprehensive Transcriptome Assembly of Chickpea (Cicer arietinum L.) Using Sanger and Next Generation Sequencing Platforms: Development and Applications. PLoS One 2014;9:e86039. https://doi.org/10.1371/JOURNAL.PONE.0086039.

[30] Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res 2011;21:1193–200. https://doi.org/10.1101/GR.113779.110.

[31] Zeng V, Villanueva KE, Ewen-Campen BS, Alwes F, Browne WE, Extavour CG. De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean Parhyale hawaiensis. BMC Genomics 2011;12:1–19. https://doi.org/10.1186/1471-2164-12-581/FIGURES/8.

[32] Azeez OI, Myburgh JG, Bosman AM, Featherston J, Sibeko-Matjilla KP, Oosthuizen MC, et al. Next generation sequencing and RNA-seq characterization of adipose tissue in the Nile crocodile (Crocodylus niloticus) in South Africa: Possible mechanism(s) of pathogenesis and pathophysiology of pansteatitis. PLoS One 2019;14:e0225073.

[33] Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, et al. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Resour 2012;12:834–45. https://doi.org/10.1111/J.1755-0998.2012.03148.X.

[34] Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nat Rev Genet 2014;2014(153):176–92. https://doi.org/10.1038/nrg3644.

[35] Ungaro A, Pech N, Martin JF, Scott McCairns RJ, Mévy JP, Chappaz R, et al. Challenges and advances for transcriptome assembly in non-model species. PLoS One 2017;12:e0185020.

[36] Freedman AH, Clamp M, Sackton TB. Error, noise and bias in de novo transcriptome assemblies. Mol Ecol Resour 2021;21:18–29. https://doi.org/10.1111/1755-0998.13156.

[37] Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for Transcript Profiling Using Long-Read Technologies. Front Genet 2020;11:606. https://doi.org/10.3389/FGENE.2020.00606/BIBTEX.

[38] Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLOS Comput Biol 2022;18:e1009730.

[39] Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics 2016;32:2210–2. https://doi.org/10.1093/bioinformatics/btw218.

[40] Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 2016;26:1134–44. https://doi.org/10.1101/gr.196469.115.

[41] Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol 2014;15:1–21. https://doi.org/10.1186/S13059-014-0553-5/TABLES/3.

[42] Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol 2018;35:543–8. https://doi.org/10.1093/molbev/msx319.

[43] Ruiz Cuevas MV, Hardy MP, Hollý J, Bonneil É, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep 2021;34. https://doi.org/10.1016/J.CELREP.2021.108815/ATTACHMENT/2D8DDBBD-780D-4DC0-9E3F-D6F3C0A8D5F6/MMC1.PDF.

[44] Lau E, Han Y, Williams DR, Thomas CT, Shrestha R, Wu JC, et al. Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. Cell Rep 2019;29:3751–3765.e5. https://doi.org/10.1016/J.CELREP.2019.11.026.

[45] Rodriguez JM, Pozo F, Di Domenico T, Vazquez J, Tress ML. An analysis of tissue-specific alternative splicing at the protein level. PLOS Comput Biol 2020;16:e1008287.

[46] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013;41:D991–5. https://doi.org/10.1093/NAR/GKS1193.

[47] Leinonen R, Sugawara H, Shumway M. Collaboration on behalf of the INSD. The Sequence Read Archive. Nucleic Acids Res 2011;39:D19–21. https://doi.org/10.1093/NAR/GKQ1019.

[48] Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 2019;47:D442–50. https://doi.org/10.1093/nar/gky1106.

[49] di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;2017(354):316–9. https://doi.org/10.1038/nbt.3820.

[50] Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 2012;28:2520–2. https://doi.org/10.1093/bioinformatics/bts480.

[51] Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics 2012;28:1525–6. https://doi.org/10.1093/BIOINFORMATICS/BTS167.

[52] Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed July 10, 2019).

[53] GitHub - mdshw5/fastqp: Simple FASTQ quality assessment using Python n.d. https://github.com/mdshw5/fastqp (accessed June 29, 2022).

[54] GitHub - ExpressionAnalysis/ea-utils: Automatically exported from code.google.com/p/ea-utils n.d. https://github.com/ExpressionAnalysis/ea-utils (accessed June 29, 2022).

[55] Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina {RNA}-seq reads. GigaScience 2015;4:48. https://doi.org/10.1186/s13742-015-0089-y.

[56] Marçais G, Yorke JA, Zimin A. QuorUM: An Error Corrector for Illumina Reads. PLoS One 2015;10:e0130821.

[57] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJournal 2011;17:10. https://doi.org/10.14806/ej.17.1.200.

[58] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

[59] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.

[60] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/BIOINFORMATICS/BTP324.

[61] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;2019(378):907–15. https://doi.org/10.1038/s41587-019-0201-4.

[62] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14:1–13. https://doi.org/10.1186/GB-2013-14-4-R36/FIGURES/6.

[63] Schiex T, Gouzy J, Moisan A, de Oliveira Y. FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nucleic Acids Res 2003;31:3738. https://doi.org/10.1093/NAR/GKG610.

[64] Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 2001;29:2607–18. https://doi.org/10.1093/NAR/29.12.2607.

[65] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008;26:1367–72. https://doi.org/10.1038/nbt.1511.

[66] Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass

spectrometry–based proteomics. Nat Methods 2017;2017(145):513–20. https://doi.org/10.1038/nmeth.4256.
[67] Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 2014;2014(51):1–10. https://doi.org/10.1038/ncomms6277.
[68] Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep 2017;18:762–76. https://doi.org/10.1016/j.celrep.2016.12.063.
[69] Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 2008;36:3420–35. https://doi.org/10.1093/NAR/GKN176.
[70] Schmid R, Blaxter ML. annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinf 2008;9:1–6. https://doi.org/10.1186/1471-2105-9-180/FIGURES/2.
[71] Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R. Annocript: a flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics 2015;31:2199–201. https://doi.org/10.1093/BIOINFORMATICS/BTV106.
[72] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
[73] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2014;2014(121):59–60. https://doi.org/10.1038/nmeth.3176.
[74] Levin M, Scheibe M, Butter F. Proteotranscriptomics assisted gene annotation and spatial proteomics of Bombyx mori BmN4 cell line. BMC Genomics 2020;21. https://doi.org/10.1186/s12864-020-07088-7.
[75] Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, et al. High-quality genome assembly of the silkworm. Bombyx mori Insect Biochem Mol Biol 2019;107:53–62. https://doi.org/10.1016/J.IBMB.2019.02.002.
[76] Wang M, Wang J, Carver J, Pullman BS, Cha SW, Bandeira N. Assembling the Community-Scale Discoverable Human Proteome. CellSyst 2018;7:412–421. e5. https://doi.org/10.1016/J.CELS.2018.08.004/ATTACHMENT/FB37BC2E-728D-4183-83BC-04C64DEB2779/MMC6.PDF.
[77] Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res 2011;10:1785–93. https://doi.org/10.1021/PR101060V/SUPPL_FILE/PR101060V_SI_002.PDF.
[78] Noor Z, Ahn SB, Baker MS, Ranganathan S, Mohamedali A. Mass spectrometry–based protein identification in proteomics—a review. Brief Bioinform 2021;22:1620–38. https://doi.org/10.1093/BIB/BBZ163.
[79] Bekker-Jensen DB, Martínez-Val A, Steigerwald S, Rüther P, Fort KL, Arrey TN, et al. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. Mol Cell Proteomics 2020;19:716–29. https://doi.org/10.1074/MCP.TIR119.001906.
[80] Kawashima Y, Nagai H, Konno R, Ishikawa M, Nakajima D, Sato H, et al. Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry. J Proteome Res 2022;21:1418–27. https://doi.org/10.1021/ACS.JPROTEOME.2C00023/ASSET/IMAGES/LARGE/PR2C00023_0006.JPEG.
[81] Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol 2018;14:e8126. , https://doi.org/10.15252/MSB.20178126.
[82] Zhang F, Ge W, Ruan G, Cai X, Guo T. Data-Independent Acquisition Mass Spectrometry-Based Proteomics and Software Tools: A Glimpse in 2020. Proteomics 2020;20:1900276. https://doi.org/10.1002/PMIC.201900276.
[83] Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. Nat Methods 2018;2018(156):440–8. https://doi.org/10.1038/s41592-018-0003-5.
[84] Nice EC. The separation sciences, the front end to proteomics: An historical perspective. Biomed Chromatogr 2021;35:e4995.
[85] Wang D, Eraslan B, Wieland T, Hallström B, Hopf T, Zolg DP, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol 2019;15:e8503. https://doi.org/10.15252/MSB.20188503.
[86] Ding N, Zhang B, Ying W, Song J, Feng L, Zhang K, et al. A time-resolved proteotranscriptomics atlas of the human placenta reveals pan-cancer immunomodulators. Signal Transduct Target Ther 2020.1–3.;2020(51):5. https://doi.org/10.1038/s41392-020-00224-5.
[87] Cifani P, Dhabaria A, Chen Z, Yoshimi A, Kawaler E, Abdel-Wahab O, et al. ProteomeGenerator: A Framework for Comprehensive Proteomics Based on de Novo Transcriptome Assembly and High-Accuracy Peptide Mass Spectral Matching. J Proteome Res 2018;17:3681–92. https://doi.org/10.1021/ACS.JPROTEOME.8B00295/SUPPL_FILE/PR8B00295_SI_002.ZIP.
[88] Ma J, Saghatelian A, Shokhirev MN. The influence of transcript assembly on the proteogenomics discovery of microproteins. PLoS One 2018;13:e0194518.
[89] Kumar D, Yadav AK, Jia X, Mulvenna J, Dash D. Integrated Transcriptomic-Proteomic Analysis Using a Proteogenomic Workflow Refines Rat Genome Annotation. Mol Cell Proteomics 2016;15:329–39. https://doi.org/10.1074/mcp.M114.047126.
[90] Müller T, Boileau E, Talyan S, Kehr D, Varadi K, Busch M, et al. Updated and enhanced pig cardiac transcriptome based on long-read RNA sequencing and proteomics. J Mol Cell Cardiol 2021;150:23–31. https://doi.org/10.1016/J.YJMCC.2020.10.005.
[91] Prasad TSK, Mohanty AK, Kumar M, Sreenivasamurthy SK, Dey G, Nirujogi RS, et al. Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res 2017;27:133–44. https://doi.org/10.1101/GR.201368.115/-/DC1.
[92] Mohien CU, Colquhoun DR, Mathias DK, Gibbons JG, Armistead JS, Rodriguez MC, et al. A Bioinformatics Approach for Integrated Transcriptomic and Proteomic Comparative Analyses of Model and Non-sequenced Anopheline Vectors of Human Malaria Parasites. Mol Cell Proteomics 2013;12:120. https://doi.org/10.1074/MCP.M112.019596.
[93] Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods 2012;9:1207–11. https://doi.org/10.1038/nmeth.2227.
[94] Desgagné-Penix I, Khan MF, Schriemer DC, Cram D, Nowak J, Facchini PJ. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol 2010;10:252. https://doi.org/10.1186/1471-2229-10-252/FIGURES/6.
[95] Lang X, Li N, Li L, Zhang S. Integrated Metabolome and Transcriptome Analysis Uncovers the Role of Anthocyanin Metabolism in Michelia maudiae. Int. J Genomics 2019;2019. https://doi.org/10.1155/2019/4393905.