BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies

Joel Vizueta[1], Alejandro Sánchez-Gracia[1], and Julio Rozas[1]

[1]Universitat de Barcelona

May 5, 2020

Abstract

Gene annotation is a critical bottleneck in genomic research, especially for the comprehensive study of very large gene families in the genomes of non-model organisms. Despite the recent progress in automatic methods, state-of-the-art tools used for this task often produce inaccurate annotations, such as fused, chimeric, partial or even completely absent gene models for many family copies, errors that require considerable extra efforts to be corrected. Here we present BITACORA, a bioinformatics solution that integrates popular sequence similarity-based search tools and Perl scripts to facilitate both the curation of these inaccurate annotations and the identification of previously undetected gene family copies directly in genomic DNA sequences. We tested the performance of BITACORA in annotating the members of two chemosensory gene families with different repertoire size in seven available genome sequences, and compared its performance with that of Augustus-PPX, a tool also designed to improve automatic annotations using a sequence similarity-based approach. Despite the relatively high fragmentation of some of these drafts, BITACORA was able to improve the annotation of many members of these families and detected thousands of new chemoreceptors encoded in genome sequences. The program creates general feature format (GFF) files, with both curated and newly identified gene models, and FASTA files with the predicted proteins. These outputs can be easily integrated in genomic annotation editors, greatly facilitating subsequent manual annotation and downstream evolutionary analyses.

Introduction

The falling cost of high-throughput sequencing (HTS) technologies made them accessible to small labs, promoting a large number of genome-sequencing projects even in non-model organisms. Nevertheless, genome assembly and annotation, especially in eukaryotic genomes, still represent major limitations (Dominguez Del Angel et al., 2018). The unique genomic characteristics of many non-model organisms, often lacking pre-existing gene models (Yandell & Ence, 2012), and the absence of closely related species with well-annotated genomes, converts the annotation process in a big challenge. State-of-the-art pipelines for *de novo* genome annotation, like BRAKER1 (Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016) or MAKER2 (Holt & Yandell, 2011), allow integrating multiple evidences such as RNA-seq, EST data, gene models from other previously annotated species or *ab initio*gene predictions (using software such as GeneMark, (Lomsadze, Burns, & Borodovsky, 2014), Exonerate (Slater & Birney, 2005), GenomeThreader (Gremme, Brendel, Sparks, & Kurtz, 2005), Augustus (M. Stanke & Waack, 2003; Mario Stanke, Diekhans, Baertsch, & Haussler, 2008) or SNAP (Korf, 2004). Some of these pipelines, such as BRAKER1, will only report those gene models with evidences. However, the gene models predicted by these automatic tools are often inaccurate, particularly for gene family members. Furthermore, these predictions can be especially inaccurate for medium or low-quality assemblies, which is a quite common situation in the increasing large number of genome drafts of non-model organisms used in molecular ecology studies. The correct annotation of gene families frequently requires additional programs, such as Augustus-PPX (Keller, Kollmar, Stanke, & Waack, 2011a), or semi-automatic, and even manual approaches, that evaluate the quality of supporting data. This latter task is usually performed in genomic annotation editors, such as Apollo, which give researchers the option to work simultaneously in the same annotation project (Lee et al., 2013).

There are a number of issues affecting the quality of gene family annotations, especially for either old or fast evolving families (Yohe et al., 2019). First, new duplicates within a family usually originate by unequal crossing-over and are found in tandem arrays in the genome, being the more recent duplicates also the physically closest (Clifton et al., 2017; Vieira, Sánchez-Gracia, & Rozas, 2007). This configuration often causes local miss-assemblies that result in the incorrect or failed identification of tandem duplicated copies (i.e., it produces artifact, incomplete, or chimeric genes along a genomic region). Secondly, the identification and characterization of gene copies in medium- to large-sized families tends to be laborious, requiring data from multiple sources, including well-annotated remote homologs and hidden Markov model (HMM) profiles. Certainly, the fine and robust identification and annotation of the complete repertory of a gene family in a typical genome draft is a challenging task that requires important additional efforts, which are very tedious to perform manually.

In order to facilitate this curation task, we have developed BITACORA, a bioinformatics pipeline to assist the comprehensive annotation of gene families in genome assemblies. BITACORA requires of a structurally annotated genome (GFF and FASTA format) or a draft assembly, and a curated database with well-annotated members of the focal gene families. The program will perform comprehensive BLAST and HMMER searches (Altschul, 1997; Eddy, 2011) to identify putative candidate gene regions (already annotated, or not), combine evidences from all searches and generate new gene models. The outcome of the pipeline consists in a new structural annotation (GFF) file along with their encoded sequences. These output sequences can be directly used to conduct downstream functional or evolutionary analyses or to facilitate a fine re-annotation in genome browsers such as Apollo (Lee et al., 2013).

Methods and implementation

Input data files

BITACORA requires: i) a data file with the genome sequences (in FASTA format); ii) the associated GFF file with annotated features (either in GFF3 or GTF formats; features must include both transcript or mRNA, and CDS); iii) a data file with the predicted proteins included in the GFF (in FASTA format); and iv) a database (here referred as FPDB database) with the protein sequences of well annotated members of the gene family of interest (focal family; in FASTA format) along with its HMM profile (see Supplementary Material for a detailed description of FPDB construction). Since sequence similarity-based searches are very sensitive to the quality of the proteins in FPDB, it is important to include in this database highly curated proteins from closely related species. This is especially important for the annotation of very old or fast-evolving gene families. Also, the use of a HMM profile increases the likelihood of identifying sequences encoding new members; these profiles can be obtained from external databases (such as PFAM) or built using high quality protein alignments with the program *hmmbuild*(Finn *et al.* , 2014). Before starting the analysis, BITACORA checks whether input data files are correctly formatted; otherwise, it will suggest some format converters distributed with the program (see Troubleshooting section in Supplementary Material).

Curating existing annotations

The BITACORA workflow has three main steps (Fig. 1). The first step consists in the identification of all putative homologs of the FPDB sequences from the focal gene family that are already present in the input GFF file, and the curation of their gene models (referred hereinafter as b-curated (bitacora-curated) gene models or proteins). Specifically, the pipeline launches BLASTP and HMMER searches (Altschul, 1997; Eddy, 2011) against the proteins predicted from the features in the input GFF using the FPDB protein sequences and HMM profiles as queries; the resulting alignments are filtered for quality (i.e. BLASTP hits covering at least two-thirds of the length of query sequences or including at least the 80% of the complete protein used as a subject are retained). The results from both searches are combined into a single integrated result for every single protein (gene model). Then, BITACORA trims the original models based in these combined results (retaining only the aligned sequence) and reports new gene coordinates (b-curated models) in a new updated GFF (uGFF), fixing for example all chimeric annotations. Besides, the proteins encoded by these b-curated models are incorporated to the FPDB (updated FPDB or uFPDB), to be used in an additional search round.

Identifying new genomic regions encoding gene family members

In the second step, BITACORA uses TBLASTN to search the genome sequences for regions encoding homologs of the proteins included in the uFPDB but not annotated in the uGFF. BITACORA implements two different approaches for generating novel gene models from TBLASTN results (set with the "gemoma" parameter). For the one hand, BITACORA implements the GeMoMa tool, a homology-based gene prediction program that uses amino acid sequence and intron position conservation to reconstruct gene models from BLAST hits (Keilwagen, Hartung, & Grau, 2019; Keilwagen, Hartung, Paulini, Twardziok, & Grau, 2018; Keilwagen et al., 2016). The second approach is based on a "close proximity" strategy. Under this strategy, all independent TBLASTN hits (i.e., after merging all alignments that overlap in TBLASTN results) located in the same scaffold and separated by less than a predetermined distance (set with the "intron distance" parameter), are connected to form a unique gene model. This step intends to join all coding exons of the same gene based on the average intron length in the focal genome. We provide some scripts to estimate this average length from the input GFF (see Supplementary Material).

Finally, to avoid reporting inaccurate gene models due to artifactual gene fusions in dense gene clusters or any other possible errors (regardless of which algorithm of the abovementioned has been applied), BITACORA will check for the presence of the gene family-specific protein domain (using the HMM profile in FPDB), and only reports in the curated dataset those gene models containing the domain. In addition, all proteins are tagged with a label that indicates the number of different domains in the sequence (Ndom). This final filtering step can be relaxed using the BITACORA "genomicblastp" option, which evaluates the presence of positive hits in either HMMER, or BLASTP searches against the proteins in FPDB (see Supplementary Material for details).

Optional search round and final output

Finally, BITACORA can also be used to perform a second search round using as the input data all proteins obtained in steps 1 and 2 (sFPDB database). This additional step (step 3 in Fig 1) is especially useful for searching remote homologs undetected in the first round. The final BITACORA outcome will include 1) an updated GFF file with both b-curated and b-novel gene models. 2) All non-redundant proteins predicted from these feature annotations (in a FASTA file). 3) Two BED files, one with the coordinates of all independent TBLASTN hits found in the genome sequence, and the other with only those hits that would encode novel putative exons and, 4) all protein sequences found in all steps.

Additional features

BITACORA could be also used in the absence of either a reference genome for the target species (e.g. for transcriptomic studies; Protein mode) or a precompiled GFF (e.g. for non-annotated genomes; Genome mode); in these cases, the input should be a FASTA file with the set of predicted proteins or the genome sequences, respectively (see Supplementary Material for alternative usage modes). With BITACORA, we also distribute a series of scripts to perform some useful tasks, such as estimating intron length statistics from a GFF, converting GFF to GTF format, and retrieving all protein sequences encoded by the features of a GFF file. Furthermore, to better adjust to the particularities of each genome, BITACORA allows the user to specify the values of the most important parameters, such as the E -value for BLAST and HMMER searches, the number of threads in BLAST runs, and the algorithm to build novel gene models from TBLASN hits.

BITACORA application example

To demonstrate the performance of BITACORA in annotating gene family members in a group of genomes of different assembly quality, we present an extended report of the results in Vizueta et al., (2018). Specifically, we selected two of the arthropod chemosensory gene families, insect gustatory receptors (GR) and Niemann-Pick type C2 (NPC2) proteins (Pelosi, Iovinella, Felicioli, & Dani, 2014; Robertson, 2015) in a subset of seven of the eleven chelicerate genomes surveyed in this study (Table 1; Fig. 2). We selected these gene families since they widely differ in the number of members and protein length. Whereas the GR is a large gene family that encode seven-transmembrane receptors of about 400 amino acids long, the NPC2 have few members and encode shorter proteins (an average of about 150 amino acids); despite the different length, both gene families have a similar average number of exons per gene in the surveyed species. Furthermore, to validate the accuracy of our software in gold standard annotated genomes, we also checked the performance of BITACORA in identifying these members in the genome of*Drosophila melanogaster* .

For the analysis, we retrieved genome sequences, annotations and predicted peptides of *D. melanogaster* (r6.31, FlyBase; Adams et al., 2000), the scorpions *Centruroides sculpturatus* (bark scorpion, genome assembly version v1.0, annotation version v0.5.3; Human Genome Sequencing Center (HGSC)) and *Mesobuthus martensii* (v1.0, Scientific Data Sharing Platform Bioinformation (SDSPB)) (Cao et al., 2013); and of the spiders *Acanthoscurria geniculata* (tarantula, v1, NCBI Assembly, BGI) (Sanggaard et al., 2014), *Stegodyphus mimosarum* (African social velvet spider, v1, NCBI Assembly, BGI) (Sanggaard et al., 2014), *Latrodectus hesperus* (western black widow, v1.0, HGSC), *Parasteatoda tepidariorum* (common house spider, v1.0 Augustus 3, SpiderWeb and HGSC) (Schwager et al., 2017) and*Loxosceles reclusa* (brown recluse, v1.0, HGSC).

In addition, and with a benchmarking purpose, we compared the performance of BITACORA with Augustus PPX, a method that also uses protein profiles to improve automatic annotations of gene family members (–proteinprofile; Keller et al., 2011; Mario Stanke, Schöffmann, Morgenstern, & Waack, 2006), in annotating GR and NPC2 copies in the same seven chelicerate genomes. Strikingly, BITACORA uncovered the identification of thousands of new gene models previously undetected in chelicerates, even after applying Augustus-PPX (Table 1; see also supplementary data in Vizueta et al. 2018 to find the BITACORA curated sequences). For instance, in the bark scorpion *Centruroides sculpturatus* , the automatic annotation pipelines show 24 GR encoding sequences, while BITACORA was able to identify and annotate 1,234 genes or gene fragments, for the only 307 recovered with Augustus-PPX (Table 1; Supplementary table S1). Globally, BITACORA identified, annotated and curated 3,570 sequences encoding GR proteins across the seven chelicerate genomes (3,466 of which were absent in the available GFF for this species), while Augustus-PPX only predicted 1,638 gene models for this family (Table1; Supplementary table S1). It is largely known that this gene family evolves rapidly in arthropods, both in terms of sequence change and repertory size, encoding in the same genome very recent and distantly related receptors as well as pseudogenes. Since some of these receptors show a very restricted gene expression pattern (expressed in specialized cells and tissues involved in chemoreception), their transcripts are often missing in RNA-seq data sets, which are one of evidences used for the automatic annotation of the genomes (Joseph & Carlson, 2015; Robertson, 2015; Vizueta et al., 2017; Zhang, Zheng, Li, & Fan, 2014). This fact, together with the huge divergence that exhibit many copies (old duplication events and/or rapid evolution), are probably the causes of the low accuracy of both automatic annotation and Augustus-PPX.

The members of the NPC2 family, on the contrary, are much more conserved at the sequence level and show higher levels of gene expression in arthropods (Pelosi et al., 2014). As expected, the number of newly identified copies is much lower than in the case of GRs. Even that, BITACORA was able to detect 44 novel NPC2 encoding sequences, raising the total annotated repertoire in these species from 75 to 119 (Table 1). In this case, Augustus-PPX was able to recover 97 gene models for this gene family, which improves the performance of previous automatic annotations, but still is outperformed by BITACORA. Importantly, Augustus-PPX predicted thousands of gene models that are not real members of the focal gene family (Supplementary table S1), requiring further actions to separate gene family copies from false allocations.

Finally, both methods correctly annotated all members of the GR and NPC2 families in *D. melanogaster* genome, demonstrating the real utility of these tools in the genome drafts of non-model organisms. It is worth noting, however, that a non-negligible number of these novel identified genes in chelicerate genomes are incomplete (about 40% and 63% of the GR and NPC2 members, respectively). This feature can be partially explained by the poor genome assembly quality (indicated as the N50 and number of scaffolds), or by the low number of annotated proteins in the input GFF. Despite BITACORA can be useful under such low-quality data, it will compromise its performance in terms of complete gene models.

Discussion

Gene families are one of the most abundant and dynamic components of eukaryotic genomes. Therefore, having curated genomic data is fundamental not only to carry out comprehensive comparative or functional genomics studies on gene families, but also to understand global genome architecture and biology. During the last decades, the rapid development of sequencing technologies has enabled the large accumulation of genome sequences of non-model organisms. These projects, which often address very specific molecular ecology studies or are in the context of large comparative genomics analyses, typically rely on automatic annotation pipelines and very little efforts are devoted to curate these annotations (see Sánchez-Herrero et al., 2019; and references therein). The proteins predicted by automatic annotation tools often contain systematic errors, such as incomplete or chimeric gene models, which are especially notable in gene families given the repetitive nature of their members. Besides, since new copies commonly arise by unequal crossing-over, they are frequently found in physically close tandem arrays of similar sequences, further complicating annotations (Clifton et al., 2017; Vieira et al., 2007).

With this in mind, we have developed a bioinformatics tool that helps researchers to access these automatic annotations, extract the information of focal gene families, curate and update gene models and identify new copies from DNA sequences. Using BITACORA, gene family annotations can be really improved using both HMM profiles and iterative searches that incorporate the new variability found in previous searches. Indeed, we validated our tool by comparing its performance with a method developed to improve the annotation of gene family members matching a protein profile, Augustus-PPX (Keller et al., 2011b; Mario Stanke et al., 2006). BITACORA not only outperforms the annotations of Augustus-PPX in the two examples showed here, but also demonstrated to be more accurate in its predictions.

The estimation of gene gains and losses, and the associated birth and death rates analyses, are very sensitive to the quality of genome annotations. The example of the GR family in chelicerates demonstrates the importance of refining annotations using BITACORA. Indeed, using unsupervised annotations in low quality genome drafts of non-model organisms directly to estimate turnover rates might produce very erroneous results, not only in terms of gene counts but also in calculations biased to highly expressed and/or very recent copies. Then, BITACORA can be used to reduce considerably these errors and make more accurate and robust inferences about the age/origin of the family and of its mode of evolution.

On the other hand, the curation of both existing and new identified members of a family with BITACORA might be also crucial for further analysis on their sequence evolution. The quality of multiple sequence alignments, which are used to determine orthology groups, to obtain divergence estimates or to detect the footprint of natural selection in gene family members, is strongly compromised by the presence of badly annotated copies, including chimeras and incorrectly annotated fragments. Using BITACORA we can detect these artifacts and either fix or discard them from further analyses.

Despite its proven utility, we are aware that BITACORA does not provide perfect annotations for a gene family. The use of GeMoMa algorithm is more sensitive than the close-proximity method generating more accurate gene models, although, in the presence of assembly errors or highly fragmented genomes, this approach might fail to identify genes, and especially putative pseudogenes. In these cases, the close-proximity method could help to detect these cases and report them in final output.

Furthermore, to overcome putative gene model errors, BITACORA implements some filtering steps to determine if the predicted coding sequences are correct. The program carries out a HMMER search to identify the protein family domain in all new annotated sequences. In addition, if the HMMER search is negative, BITACORA can relax this step by checking if the novel genes show significant BLASTP hits in a search against FPDB proteins. In this case, the sensitivity of the annotations will increase at the expense of specificity (i.e. it could generate false allocations to the focal family in the presence of repetitive regions or

FPDB contaminations, for instance). It is important to note that BITACORA generates homology-based predictions that could require different levels of experimental validation depending on the nature of further downstream analyses.

Notwithstanding such filtering steps, BITACORA offers an output directly readable in genome editor tools, such as Apollo, which facilitate researchers to improve gene models. Fig. 3 shows an example of the annotation tracks generated by BITACORA (GFF3 and BED files) for a cluster of three members of the NPC2 family in the genome of the spider*P. tepidariorum* . The automatic annotation of this region using MAKER2 (track Ptep_v0.5.3-Models), generated a chimeric gene model (two different genes are fused) which could be easily curated using BITACORA. Additionally, despite TBLASTN searches detected a putative novel exon in the gene encoding NPC2_5, GeMoMa did not include this sequence in the final gene model due to the presence of an in-frame stop codon. In order to decide if this stop codon is an annotation, assembly or sequencing artifact, it would be necessary, for instance, to verify if the exon exists in other species, if that region is transcribed, or if the gene is under selective constraints.

Conclusion

Genome annotation, especially in medium to low quality drafts of non-model organisms, is still a drawback for the increasingly large number of evolutionary and functional genomic analyses in the context of molecular ecology studies. To assists this task, we developed a comprehensive pipeline that facilitates the curation of existing models and the identification of new gene family copies in genome assemblies. The improved annotations generated with this pipeline can be used directly to perform downstream analyses or as a baseline for further manual curation in genomic annotation editors. Future directions should include the possibility of including novel sources of evidence in BITACORA searches, such as RNA-seq data, or the integration of the pipeline as a part of genome annotation editors to facilitate gene family annotation in collaborative genome projects.

Acknowledgements

We would like to thank Paula Escuer and Vadim Pisarenco for helpful discussions. This work was supported by the Ministerio de Economía y Competitividad of Spain (CGL2013-45211, CGL2016-75255) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2017SGR1287). J.V. was supported by a FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437).

Author contributions

J.V., A.S.-G and J.R. conceived the work. J.V. wrote the scripts, did the analyses and wrote the first version of the manuscript. All authors checked and confirmed the final version of the manuscript.

Data accessibility

BITACORA is available from http://www.ub.edu/softevol/bitacora, and https://github.com/molevol-ub/bitacora

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster* . *Science* ,*287* (5461), 2185–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10731132

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* ,*25* (17), 3389–3402. doi:10.1093/nar/25.17.3389

Clifton, B. D., Librado, P., Yeh, S.-D., Solares, E. S., Real, D. A., Jayasekera, S. U., … Ranz, J. M. (2017). Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in *Drosophila* . *Molecular Biology and Evolution* , *34* (1), 51–65. doi:10.1093/molbev/msw212

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., … Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation.*F1000Research* , *7* , ELIXIR-148. doi:10.12688/f1000research.13598.1

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology* , *7* (10), e1002195. doi:10.1371/journal.pcbi.1002195

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research* , *42* (Database issue), D222–D230. doi:10.1093/nar/gkt1223

Gremme, G., Brendel, V., Sparks, M. E., & Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms.*Information and Software Technology* , *47* (15), 965–978. doi:10.1016/J.INFSOF.2005.09.005

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* , *32* (5), 767–769. doi:10.1093/bioinformatics/btv661

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.*BMC Bioinformatics* , *12* (1), 491. doi:10.1186/1471-2105-12-491

Joseph, R. M., & Carlson, J. R. (2015). *Drosophila*Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain. *Trends in Genetics : TIG* , *31* (12), 683–695. doi:10.1016/j.tig.2015.09.005

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. In *Methods in Molecular Biology* (Vol. 1962, pp. 161–177). Humana Press Inc. doi:10.1007/978-1-4939-9173-0_9

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* , *19* (1), 189. doi:10.1186/s12859-018-2203-5

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* ,*44* (9), 89. doi:10.1093/nar/gkw092

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011a). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* , *27* (6), 757–763. doi:10.1093/bioinformatics/btr010

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011b). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* , *27* (6), 757–763. doi:10.1093/bioinformatics/btr010

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* , *5* , 59. doi:10.1186/1471-2105-5-59

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., … Lewis, S. E. (2013). Web Apollo: a web-based genomic annotation editing platform. *Genome Biology* ,*14* (8), R93. doi:10.1186/gb-2013-14-8-r93

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* , *42* (15), e119–e119. doi:10.1093/nar/gku557

Pelosi, P., Iovinella, I., Felicioli, A., & Dani, F. R. (2014). Soluble proteins of chemical communication: an overview across arthropods.*Frontiers in Physiology* , *5* (August), 320. doi:10.3389/fphys.2014.00320

Robertson, H. M. (2015). The Insect Chemoreceptor Superfamily Is Ancient in Animals. *Chemical Senses* , *40* (9), 609–614. doi:10.1093/chemse/bjv046

Sanchez-Herrero, J. F., Frias-Lopez, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sanchez-Gracia, A., & Rozas, J. (2019). The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. *GigaScience 8* (8), 1-9. doi:10.1093/gigascience/giz099

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* , *6* , 31. doi:10.1186/1471-2105-6-31

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* , *19* (Suppl 2), ii215–ii225. doi:10.1093/bioinformatics/btg1080

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* , *24* (5), 637–644. doi:10.1093/bioinformatics/btn013

Stanke, M., Schoffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* ,*7* (1), 62. doi:10.1186/1471-2105-7-62

Vieira, F. G., Sanchez-Gracia, A., & Rozas, J. (2007). Comparative genomic analysis of the odorant-binding protein family in 12*Drosophila* genomes: purifying selection and birth-and-death evolution. *Genome Biology* , *8* (11), R235. doi:10.1186/gb-2007-8-11-r235

Vizueta, J., Frias-Lopez, C., Macias-Hernandez, N., Arnedo, M. A., Sanchez-Gracia, A., & Rozas, J. (2017). Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biology and Evolution* , *9* (1), 178–196. doi:10.1093/gbe/evw296

Vizueta, J., Rozas, J., & Sanchez-Gracia, A. (2018). Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates. *Genome Biology and Evolution* , *10* (5), 1221–1236. doi:10.1093/gbe/evy081

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* , *13* (5), 329–342. doi:10.1038/nrg3174

Yohe, L. R., Davies, K. T. J., Simmons, N. B., Sears, K. E., Dumont, E. R., Rossiter, S. J., & Davalos, L. M. (2019). Evaluating the performance of targeted sequence capture, RNA-Seq, and degenerate-primer PCR cloning for sequencing the largest mammalian multigene family.*Molecular Ecology Resources* . doi:10.1111/1755-0998.13093

Zhang, Y., Zheng, Y., Li, D., & Fan, Y. (2014). Transcriptomics and identification of the chemoreceptor superfamily of the pupal parasitoid of the oriental fruit fly, *Spalangia endius* Walker (Hymenoptera: Pteromalidae). *PloS One* , *9* (2), e87800. doi:10.1371/journal.pone.0087800

Tables

Table 1. Summary of the number of GRs and NPC2 genes identified by BITACORA and Augustus-PPX in genome assemblies.

Figures

Fig. 1. Schematic representation of the BITACORA workflow.

Fig. 2. Phylogenetic relationships among the seven chelicerate species surveyed for the GR and the NPC2 families.

Fig. 3. Example of the visualization in the Apollo genome editor of the BITACORA output. The example includes the annotation features of three genes encoding NPC2 proteins that are arranged in tandem in the spider *P. tepidariorum* . Current automatic annotation of this genomic region obtained with MAKER2

(track PTEP_v0.5.3-Models), produced a chimeric gene model (PtepTmpM024154-RA; an artifactual two genes fusion), which is effectively curated by BITACORA (NPC2_5 and NPC2_6 gene models). The next three tracks are generated by BITACORA. The GFF3_NPC2_BITACORA track, which includes the final gene models, both curated or newly identified by the program, and the BED_NPC2_All and BED_NPC2_Novel tracks showing the position of all independent TBLASTN hits found in sequence similarity-based searches, or only those involving novel putative exons, respectively. Note that a novel coding sequence (not predicted in automatic annotations) is predicted by the program.

Supplementary Material

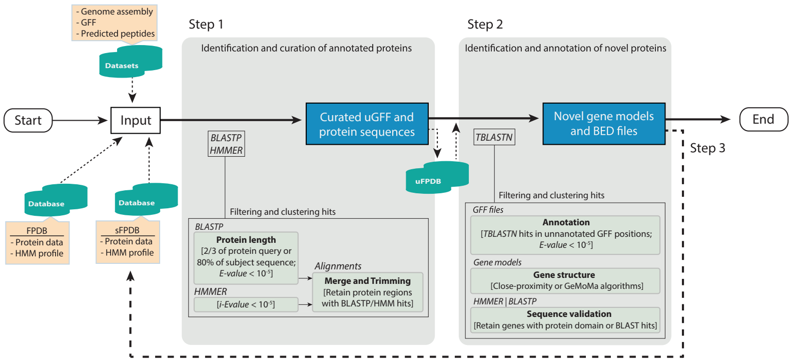
Table S1. Summary of the genome information and the number of GRs and NPC2 genes identified by BITACORA and Augustus-PPX in the genome assemblies of the seven surveyed chelicerates, and in *D. melanogaster* .

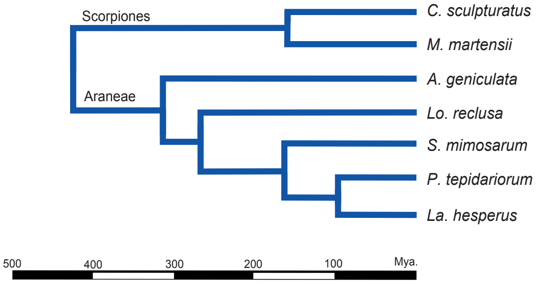
Supplementary documentation

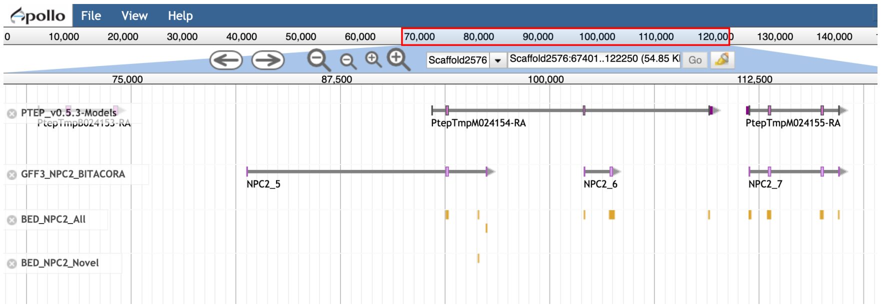
BITACORA Documentation

Hosted file

`Table1_bitacora_12Mar20.xlsx` available at https://authorea.com/users/304673/articles/435223-bitacora-a-comprehensive-tool-for-the-identification-and-annotation-of-gene-families-in-genome-assemblies


- Genome assembly
- GFF
- Predicted peptides
Datasets
Step 1
Identification and curation of annotated proteins
Step 2
Identification and annotation of novel proteins
Start
Input
Curated uGFF and protein sequences
Novel gene models and BED files
End
Step 3
BLASTP
HMMER
uFPDB
TBLASTN
Database
FPDB
- Protein data
- HMM profile
Database
sFPDB
- Protein data
- HMM profile
Filtering and clustering hits
BLASTP
Protein length
[2/3 of protein query or 80% of subject sequence; E-value < 10-5]
HMMER
[i-Evalue < 10-5]
Alignments
Merge and Trimming
[Retain protein regions with BLASTP/HMM hits]
Filtering and clustering hits
GFF files
Annotation
[TBLASTN hits in unnanotated GFF positions; E-value < 10-5]
Gene models
Gene structure
[Close-proximity or GeMoMa algorithms]
HMMER | BLASTP
Sequence validation
[Retain genes with protein domain or BLAST hits]


Scorpiones
Araneae
C. sculpturatus
M. martensii
A. geniculata
Lo. reclusa
S. mimosarum
P. tepidariorum
La. hesperus
500
400
300
200
100
Mya.


Apollo
File
View
Help
0
10,000
20,000
30,000
40,000
50,000
60,000
70,000
80,000
90,000
100,000
110,000
120,000
130,000
140,000
Scaffold2576
Scaffold2576:67401..122250 (54.85 K
Go
75,000
87,500
100,000
112,500
PTEP_v0.5.3-Models
PtepTmpB024153-RA
PtepTmpM024154-RA
PtepTmpM024155-RA
GFF3_NPC2_BITACORA
NPC2_5
NPC2_6
NPC2_7
BED_NPC2_All
BED_NPC2_Novel