Nov 18, 2025 Version 1

Structural Genome Annotation QC with GAQET2 V.1

DOI
https://dx.doi.org/10.17504/protocols.io.8epv5k5rnv1b/v1

Victor Garcia-Carpintero[1], Iñigo de Martín[1], Sofía García-Juan[1], Aureliano Bombarely[1]

[1]Institute of Molecular and Celular Plant Biology (IBMCP)

Victor Garcia-Carpintero

DOI: https://dx.doi.org/10.17504/protocols.io.8epv5k5rnv1b/v1

External link: https://github.com/victorgcb1987/GAQET2

Protocol Citation: Victor Garcia-Carpintero, Iñigo de Martín, Sofía García-Juan, Aureliano Bombarely 2025. Structural Genome Annotation QC with GAQET2. **protocols.io** https://dx.doi.org/10.17504/protocols.io.8epv5k5rnv1b/v1Version created by **Aureliano Bombarely**

Protocol status: Working
This protocol was successfully followed by external students.

Created: October 30, 2025

Last Modified: November 18, 2025

Protocol Integer ID: 231154

Keywords: Genomics, Bioinformatics, Annotation, Benchmarking, Quality Control, structural genome annotation qc with gaqet2, quality of the genome structural annotation, thee quality of the gene model annotation, structural genome annotation qc, genome structural annotation, gene model annotation, genome annotation process, target species genome, identification of gene model, phylogenetic distance of the protein data, gene model, sequence homology search, sequence homology searches by diamond, sequencing technology, phylogenetic distance, hmw dna extraction, gaqet2, gene, gene space, protein data, fast development of the dna, dna, tools like agat

Funders Acknowledgements:
Agencia Española de Investigación (AEI)
Grant ID: PID2022-142508NB-I00

Disclaimer

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to **protocols.io** is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with **protocols.io**, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Abstract

The fast development of the DNA sequencing technologies has moved the bottleneck to produce high quality genomes to the HMW DNA extraction and to the genome annotation processes. The identification of gene models can be performed based on experimental evidence and/or ab-initio models. Many factors such as incomplete representation of the gene space, phylogenetic distance of the protein data to the target species genome, specificity of the ab-initio models, and transposon dynamics can impact negatively the quality of the genome structural annotation. GAQET2 is a tool designed to assess thee quality of the gene model annotation for non-model species, combining tools like AGAT, BUSCO, OMArk, PSAURON, and sequence homology searches by Diamond. It also has a module to detect TE miss-identified as genes called DeTEnGA.

Image Attribution

Image created by Sofia García Juan

Guidelines

1.- Get one or more genome assemblies and their associated Structural genome annotation. In case that the annotation doesn't exists, create one first.
2.- Run GAQET2 with all the analyses provided by the program and get asummary of annotation's quality metrics.
3.- In case of more than one annotation is available, redo steps 2 and 3 and generate a GAQET plot to compare all the annotations at once.
4.- Review manually summaries generated with GAQET2 to explore more specific details about your annotation(s).

Materials

The following files are needed to assess SGA's quality:

1. **Genome assembly**: this file should be in FASTA format
2. **SGA's file:** an SGA annotation file associated to the genome assembly. Supported formats are GTF and GFF3
3. **Proteome file** (optional): this files are only needed if you are going to run homology seach analysis. Multiple proteome files in FASTA format can be included. We recommend using TrEMBL and SWISSPROT and can be found at **https://www.uniprot.org/**. included and for running OMArk and Orhtologous MAtrix (OMA) database file is needed.
4. **Ortholog MAtrix (OMA) database file** (optional): This file is only need for OMArk's taxonomic consistency and Completness analyses. It can be downloaded from the OMArk's github webpage or using the following link https://omabrowser.org/All/LUCA.h5.
5. **YAML configuration file**: This file is a configuration file needed to run GAQET

Troubleshooting

Before start

GAQET2 runs a variety of different programs to generate the Structural Genome Annotation (SGA) quality benchmarking. The following section explains GAQET2 installation.

GAQET2 installation guide

- InterproScan

We recommend using the github version instead of any conda installation (https://github.com/ebi-pf-team/interproscan-docs/blob/v5/docs/HowToDownload.rst) Then, add interproscan.sh to your PATH variable:

```
export PATH=$PATH:/path/to/interproscan.sh
```

- Other programs, including GAQET2 itself

GAQET2 and all other dependencies are included in a conda package. We recommend the miniconda version to minimize the disk space needed and reduce unnecesary software download. Following links has download and installation details for miniconda: https://www.anaconda.com/docs/getting-started/miniconda/main. Then, GAQET2 can be installed by executing the following commands:

```
conda create -n GAQET
conda activate GAQET
conda install -c victorgcb gaqet
```

We can check if installation was succesfuly by typing in the command line:

```
GAQET -v
```

Structural Genome Annotation metrics generation

1 **Preparing input files**

The following files are needed to assess SGA's quality:

1. **Genome assembly**: this file should be in FASTA format
2. **SGA's file:** an SGA annotation file associated to the genome assembly. Supported formats are GTF and GFF3
3. **Proteome file** (optional): this files are only needed if you are going to run homology search analysis. Multiple proteome files in FASTA format can be included. We recommend using TrEMBL and SWISSPROT (1) and can be found at **https://www.uniprot.org/**.
4. **Ortholog MAtrix (OMA) database file** (optional): This file is only need for OMArk's (2) taxonomic consistency and Completness analyses. It can be downloaded from the OMArk's github webpage or using the following link https://omabrowser.org/All/LUCA.h5.
5. **YAML configuration file**: This file is a configuration file needed to run GAQET. A typical YAML configuration for GAQET has the following structure:

```
ID: "SpeciesName"
Assembly: "/path/to/assembly.fasta"
Annotation: "/path/to/annotation.gff3"
Basedir: "/path/to/GAQET/results"
Threads: 20
Analysis:
  - AGAT
  - BUSCO
  - PSAURON
  - DETENGA
  - OMARK
  - PROTHOMOLOGY
OMARK_db: "/path/to/omark_db.h5"
OMARK_taxid: NCBItaxonID
BUSCO_lineages:
  -  clade1_odb10
  -  clade2_odb10
PROTHOMOLOGY_tags:
  - TREMBL: "/path/to/uniprot_trembl_db.dmnd"
  - SWISSPROT: "/path/to/uniprot_swssprot.dmnd"
DETENGA_db: "rexdb-plant"
```

The following table explains each YAML configuration feature:

	A	B
	Parameter	**Description**
	DETENGA_db	DeTEnGA database for interpro checks. Only needed if DETENGA is in Analysis
	Assembly	FASTA genome file
	Basedir	GAQET2 Results directory
	Annotation	GFF3/GTF annotation file
	Analysis	List of analysis to run. All of them are optional
	BUSCO_lineages	List of BUSCO clades to run. Only needed if BUSCO is in Analysis. One or more can be included
	PROTHOMOLOGY_tags	List of name and path to DIAMOND proteins database. Only needed if PROTHOMOLOGY is in Analysis. One or more can be included. Tag Names are user-defined
	ID	Name of the species
	OMARK_taxid	NCBI taxid for OMArk. Only needed if OMArk is in Analysis
	Threads	Number of threads
	OMARK_db	Path to omark db. Only needed if OMArk is in Analysis
	DETENGA_db	DeTEnGA database for interpro checks. Only needed if DETENGA is in Analysis

- NCBI taxid can be found at NCBI's webpage. For example, *Arabidopsis thaliana* taxid is 3702 (https://www.ncbi.nlm.nih.gov/search/all/?term=arabidopsis%20thaliana, Taxonomic ID found in the Taxonomy description section).
- BUSCO (3) lineages identifiers can be found at https://busco.ezlab.org/list_of_lineages.html. Select one or more of the nearest clades included in this list. For example, for *Arabidopsis thaliana* I would use "viridiplantae_odb10" and "embryophyta_odb10" (**note that the date part of the name should not be included**).
- Detenga_db is a list of PFAMs (4) related to transposable elements found in the rexdb database. Choices are "rexdb-plant", "rexdb-metazoa" and "rexdb" for plants, metazoa and all combined organisms respectively.

2 **Running the program**

The simplest way to run the program is:

```
GAQET --YAML
```

GAQET generate an output base dir by default named "AnnotationQC_{currentdate}". If YAML configuration is correctly filled this should be enough to run the benchmarking analyses. GAQET includes optional arguments that helps to override YAML configuration options:

	A	B
	Parameter	**Description**
	--species, -s	Override YAML species ID
	--genome, -g	Override YAML Assembly
	--annotation, -a	Override YAML Annotation
	--taxid, -t	Override NCBI taxid
	--outbase, -o	Override YAML outbase

For example, imagine that you want to run GAQET2 with 3 plant species, *Arabidopsis thaliana*, *Vitis vinifera* and *Oryza sativa*. You want to run the same analyses using the same databases, but of course each species have their own genome assembly, annotation file and NCBI's taxon ID, and probably you want output directories with custom names for each species. In order to facilitate the running of these three species, you can prepare a YAML configuration file like this one:

```
ID: "Whatever You Like, this is going to be overriden"
Assembly: "Whatever You Like, this is going to be overriden"
Annotation: "Whatever You Like, this is going to be overriden"
Basedir: "Whatever You Like, this is going to be overriden"
Threads: 20
Analysis:
  - AGAT
  - BUSCO
  - PSAURON
  - DETENGA
  - OMARK
  - PROTHOMOLOGY
OMARK_db: "/path/to/omark_db.h5"
OMARK_taxid: Whatever You Like, this is going to be overriden
BUSCO_lineages:
  -  embryophyta_odb10
  -  viridiplantae_odb10
PROTHOMOLOGY_tags:
  - TREMBL: "/path/to/uniprot_trembl_db.dmnd"
  - SWISSPROT: "/path/to/uniprot_swssprot.dmnd"
DETENGA_db: "rexdb-plant"
```

Let's say that you have called this YAML file "GAQET_PLANTS.yaml". Then, you can run GAQET for each one of this species reusing this file and override their values like this:

```
#Arabidopsis thaliana
GAQET -i GAQET_PLANTS.yaml -s Arabidopsis_thaliana -o
Arabidopsis_thaliana_results -g Arabidopsis_thaliana.genomic.fna -
a Arabidopsis.annotation.gff -t 3702

#Vitis vinifera
GAQET -i GAQET_PLANTS.yaml -s Vitis_vinifera -o
Vitis_vinifera_results -g Vitis_vinifera.genomic.fna -a
Vitis_vinifera.annotation.gff -t 29760

# Oryza sativa
GAQET -i GAQET_PLANTS.yaml -s Oryza_sativa -o Oryza_sativa_results
-g Oryza_sativa.genomic.fna -a Oryza_sativa.annotation.gff -t 4530
```

Output folder

Each GAQET2 run should generate an output folder with this structure:

```
📁 outpudir
├── 📁 input_sequences
├── 📁 AGAT_run
├── 📁 BUSCOCompleteness_run
├── 📁 DETENGA_run
├── 📁 DIAMOND_run
├── 📁 OMARK_run
├── 📁 PSAURON_run
├── 📄 GAQET.log.txt
├── 📄 {species}_GAQET.stats.tsv
```

Each of the "{analysis}_run" directories contains the output of each analysis run. The "GAQET.log.txt" file contains things like run errors or time consumed running analysis. All programs outputs are parsed and their results are stored in a tsv file, the " {species}_GAQET.stats.tsv" file.

The log file

While GAQET2 is running, a log will be printed on screen and also is written into the GAQET.log.txt file. There is as description of how this file looks like:

```
                          ###############
                          ##      GAQET      ##
                          ###############

v1.11.17
Tue Oct 28 12:19:12 2025

-----Checking if all required inputs are present-----
        ✓       All required inputs are present
-----Checking if all analysis are valid-----
        ✓       All analysis are valid
-----Checking if BUSCO lineages are valid-----
        ✓       BUSCO lineages are valid
-----Checking if OMARK taxid is valid-----
        ✓       Taxid for OMARK is valid
-----Checking if OMARK db is available
        ✓       OMARK_db /data/shared_dbs/omark/LUCA.h5 found
-----Checking if DeTEnGA db is available
        ✓       DETENGA_db rexdb-plant found
-----Checking if protein databases exists-----
        ✓       All protein databases are valid

You can redo this analysis using the following command:

CMMD: gaqet.py -i config.yaml -s Arabidopsis_thaliana_ANv1 -t 3702
-g Athaliana_447_TAIR10.fa   -a
Athaliana_447_TAIR10.gene_models_ANv1.gff3 -o test

-----Checking binaries for gffread-----
        ✓       Binary gffread found

-----Checking binaries for seqtk-----
        ✓       Binary seqtk found

-----Checking binaries for AGAT-----
        ✓       Binary agat_sp_statistics.pl found
        ✓       Binary agat_sp_flag_premature_stop_codons.pl found
        ✓       Binary
agat_sp_filter_incomplete_gene_coding_models.pl found

-----Checking binaries for BUSCO-----
        ✓       Binary busco found

-----Checking binaries for PSAURON-----
```

```
        ✓       Binary psauron found

-----Checking binaries for DETENGA-----
        ✓       Binary TEsorter found
        ✓       Binary interproscan.sh found

-----Checking binaries for OMARK-----
        ✓       Binary omamer found
        ✓       Binary omark found

-----Checking binaries for PROTHOMOLOGY-----
        ✓       Binary diamond found
```

The first part is a print of the command used (CMMD line) an an **initial check of binaries availability** (if all programs needed to run analyses are found). GAQET will stop if one or more binaries are not found.

Log file also includes information about **sequence and annotations reformatings**:

```
-----Reformatting transcript features to mRNA-----

#Changed transcript features to mRNA
The following transcripts have been changed:
 ID=ChrC-g13.t1;Parent=ChrC-g13
ID=ChrC-g14.t1;Parent=ChrC-g14
ID=ChrC-g15.t1;Parent=ChrC-g15
ID=ChrC-g16.t1;Parent=ChrC-g16
ID=ChrC-g17.t1;Parent=ChrC-g17

-----Splitting annotation by features-----

#Separate annotation by type, command used:
        agat_sp_separate_by_record_type.pl  --gff
test/reformatted_annotation/reformatted_annotation.gff3 -o
test/input_sequences

        ✓       AGAT separate by type run successfully

Time consumed getting  splitting annotation: 83.39s

-----Checking if Assembly file has a correct length format-----

-----Getting longest isoforms-----

#Getting longest isoforms, command used:
        agat_sp_keep_longest_isoform.pl -gff
test/input_sequences/mrna.gff -o
test/input_sequences/Athaliana_447_TAIR10.longest_isoform.gff3

        ✓       AGAT longest isoform run successfully

Time consumed getting longest isoforms: 86.52s

-----Extracting CDS and protein sequences-----

#cds extraction, command used:
        gffread -x
test/input_sequences/Athaliana_447_TAIR10.cds.fasta -J -g
Athaliana_447_TAIR10.fa test/input_sequences/mrna.gff

        ✓       GFFread, mode cds run successfully

#proteins extraction, command used:
        gffread -y
test/input_sequences/Athaliana_447_TAIR10.proteins.fasta -J -g
Athaliana_447_TAIR10.fa test/input_sequences/mrna.gff

        ✓       GFFread, mode proteins run successfully

#mrna extraction, command used:
        gffread -w
test/input_sequences/Athaliana_447_TAIR10.mrna.fasta -J -g
Athaliana_447_TAIR10.fa test/input_sequences/mrna.gff
```

It also includes **analyses information** about if it have worked properly (and if not, the error produced by the analysis program will be shown) command used and time consumed:

```
-----Running AGAT on the GFF file-----

#AGAT stats command used:
        agat_sp_statistics.pl --gff test/input_sequences/mrna.gff
-o test/AGAT_run/Arabidopsis_thaliana_ANv1.01_agat_stats.txt

        ✓       AGAT stats run successfully

#AGAT stop codons command used:
        agat_sp_flag_premature_stop_codons.pl --gff
test/input_sequences/mrna.gff --fasta Athaliana_447_TAIR10.fa -o
test/AGAT_run/Arabidopsis_thaliana_ANv1.01_agat_premature_stop.txt

        ✓       AGAT premature stop codons analysis run
successfully

#AGAT incomplete CDS command used:
        agat_sp_filter_incomplete_gene_coding_models.pl --add_flag
--gff test/input_sequences/mrna.gff --fasta
Athaliana_447_TAIR10.fa -o
test/AGAT_run/Arabidopsis_thaliana_ANv1.01_agat_incomplete.txt

        ✓       AGAT incomplete CDS analysis run successfully

Time consumed Running AGAT: 294.06s

-----Running BUSCO-----

#viridiplantae_odb10 command used:
        busco --cpu 48 -i
test/input_sequences/Athaliana_447_TAIR10.proteins_longest_busco.r
enamed.fasta -o test/BUSCOCompleteness_run/viridiplantae_odb10 -m
prot -l viridiplantae_odb10 --force --tar

        ✓       BUSCO analysis with lineage viridiplantae_odb10
run successfully

#embryophyta_odb10 command used:
        busco --cpu 48 -i
test/input_sequences/Athaliana_447_TAIR10.proteins_longest_busco.r
enamed.fasta -o test/BUSCOCompleteness_run/embryophyta_odb10 -m
prot -l embryophyta_odb10 --force --tar
```

```
          ✓       BUSCO analysis with lineage embryophyta_odb10 run
successfully

Time consumed Running BUSCO: 152.65s
```

Annotation metrics explanation

3 **General metrics**

	Parameter	Description
	Species	Value assigned by the user when running the command
	NCBI_TaxID	Value assigned by the user when running the command
	Assembly_Version	Actually, the name of the assembly file. GAQET_REVIEWER provides md5sums of this file for better identification
	Annotation_Version	Actually, the name of the annotation file. GAQET_REVIEWER provides md5sums of this file for better identification
	Gene_Models (N)	(AGAT) Number of coding genes found in the annotation
	Transcript_Models (N)	(AGAT) Number of coding transcripts found in the annotation
	CDS_Models (N)	(AGAT) Number of CDS found in the annotation
	UTR5' (N)	(AGAT) Number of UTR5' annotated on coding transcripts
	UTR3' (N)	(AGAT) Number of UTR3' annotated on coding transcripts
	Both sides UTR' (N)	(AGAT) Number of coding transcripts models with both UTR's annotated
	Overlapping_Gene_Models (N)	(AGAT) Number of overlapping coding genes

	Parameter	Description
	Single Exon Gene Models (N)	(AGAT) Number of monoexonic coding genes
	Single Exon Transcripts (N)	(AGAT) Number of monoexonic coding transcripts
	Total Gene Space (Mb)	(AGAT) Coding sequences' total size
	Mean Gene Model Length (bp)	(AGAT) Average coding gene length
	Mean CDS Model Length (bp)	(AGAT) Average CDS length
	Mean Exon Length (bp)	(AGAT) Average coding exon length
	Mean Intron Length (bp)	(AGAT) Average coding gene's intron length
	Longest Gene Model Length (bp)	(AGAT) Longest coding gene length
	Longest CDS Model Length (bp)	(AGAT) Longest CDS length
	Longest Intron Length (bp))	(AGAT) Longest coding gene's intron length
	Shortest Gene Model Length (bp)	(AGAT) Shortest coding gene length
	Shortest CDS Length (bp)	(AGAT) Shortest CDS length
	Shortest intron Length (bp)	(AGAT) Shortest intron length
	Models with early STOP (N)	(AGAT) Number of coding transcripts with premature stop codons
	Models START missing	(AGAT) Number of coding transcripts lacking start codon

Parameter	Description
Models START & STOP missing	(AGAT) Number of coding transcripts lacking stop and start codon
Annotation_BUSCO_{DB}	(BUSCO) Busco proteome completness for the database {DB}. Refer to https://busco.ezlab.org/busco_userguide.html#interpreting-the-results to get an output's explanation
PSAURON SCORE	(PSAURON) Global Annotation's accuracy at detecting ORFs
DETENGA_FPV	(DeTEnGA) Number of classified transcripts. See table below for nomeclature explanation
DETENGA_FP%	(DeTEnGA) Classified transcripts in percentages. See table below for nomeclature explanation
OMArk Consistency Results	(OMARK) Taxonomic consistency results. Check table below for nomenclature explanation
OMArk Completeness Results	(OMARK) Taxonomic Completness results. Check table below for nomenclature explanation
OMArk Species Composition	(OMARK) Species Composition in percentage
ProteinsWith{db}Hits (%)	(DIAMOND) Percentage of proteins with a significant hit on database {db}

4 **BUSCO metrics**

Check https://busco.ezlab.org/busco_userguide.html#interpreting-the-results for more information.

A	B
Nomenclature	Description
C	% of orhologs found, complete set secuence

	A	B
	S	% of orthologs found, complete sequence and not duplicated
	D	% of orthologs found, complete sequence and duplicated
	F	% of orthogs found, sequence incomplete
	M	% of orthogs missing in our proteome

5 **OMArk Consistency metrics explanation**

Consistency is a quality measurement and describes the proportion of protein sequences placed into known gene families from the same lineage. Check (2) for an in-depth explanation.

	Nomeclature	Description
	Cons	Taxonomic consistent hits (%)
	Inco	Taxonomic inconsistent hits (%)
	Cont	Contaminantion hits (%)
	Unkn	Unkown hits (%)
	P	Partial hits (%)
	F	Fragmented hits (%)

6 **OMArk Completness metrics explanation**

Completeness describes how our proteome overlaps with a conserved ancestral gene set of the species' lineage. The gene set is classified in hierarchical orthologous groups (**HOGs**). Each HOG represents a single ancestral gene. Check (2) for a in-depth explanation.

	Nomeclature	Description
	{Taxon}-HOGs	Number of HOGs in our species' nearest {taxon}

	Nomeclature	Description
	S	HOGs hits by a single query protein (%)
	D	HOGs hits by more than one query protein (%)
	U	HOGs hits by more than one query protein, unexpected (no HOG duplication evidence exists)(%)
	E	HOGs hits by more than one query protein, expected (HOG duplication evidence exists, known HOG subfamilies)(%)
	M	HOGs without hit (%)

Annotation metrics explanation

7 **DeTEnGA metrics explanation**

Detection of Transposable Elements as Genes on Annotations (DeTEnGA) is an in-house tool created for coding sequences **classification as a Transposable element (TE) or not**. This classification is at protein sequence level using interpro and at mRNA level using TEsorter. Table below describes the nomenclature used in DeTEnGA:

	Nomeclature	Description
	T	Total number of coding transcripts (includes all classifications)
	PcpM0	Transcripts with non-TEs interpro PFAMs and non-TE mRNA
	PTeM0	Transcripts TEs only interpro PFAMs and non-TE mRNA
	PchM0	Transcripts with mixed TEs and non-TEs pfams and non-TE mRNA
	PcpMTe	Transcripts with non-TEs interpro PFAMs and TE mRNA
	PteMte	Transcripts with only TEs interpro PFAMs and TE mRNA
	PchMte	Transcripts with mixed TEs and non-TEs pfams and non-TE mRNA

GAQET plot generation and interpretation

8 **GAQET plot**

GAQET2 include a visual representation of the most important SGA's QC metrics generated while running the program. This plot can be generated with one run or more and can be used to classify each annotation by their overall quality in an intuitive way.

File preparation

You need at least one GAQET2 annotation summary file (the "{species}_GAQET.stats.tsv" file). For visualizing more than one summary file, a merged file should be prepared as the one shown in the following example:

FILE Arabidopsis_thaliana_AllAnnotation... 8KB

It has the same format than a single stats file but has added rows for each one of the summaries to compare.

GAQET plot generation and interpretation

9 **Running GAQET Plot**

Running GAQET plot only need the metrics summary file and an plot output filename. Filename suffix will be used for deciding in which format the plot will be encoded (".png", ".svg")

```
GAQET_PLOT -i
Arabidopsis_thaliana_AllAnnotationsGAQET20251028.stats.tsv -o
results.png
```

GAQET plot generation and interpretation

10 **GAQET plot example**

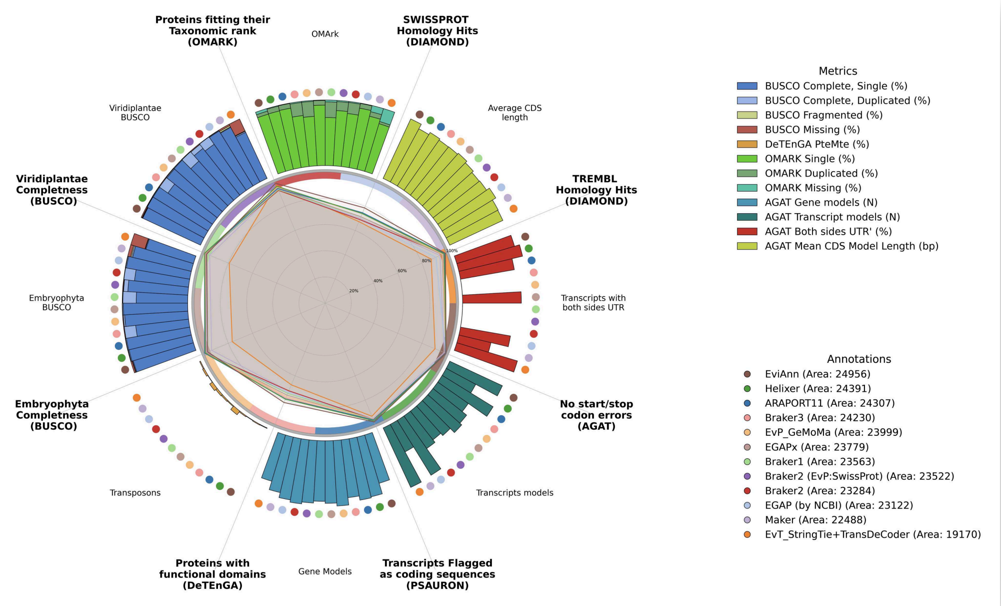

Figure 1. GAQET plot comparison. The internal radar plot shows global percentage metrics. An area score is calculated using these values as a global score of the annotation outcome (values shown in Annotations, higher means better). External bar plots complements the internal radar plot, adding detailed information about completeness, potential transposable elements and model metrics. Each kind of annotation is marked by the color circles and acronyms shows kind of methodology and program used. **EVT**: transcriptomic alignments, **EVP**: protein alignments.

This plot has 3 levels:

- **Inner radar plot**: Represents percentages for each of the metrics marked by the radial axis, for example, "No start/stop codons". Higher percentages mean better, but the outer column sets and the GAQET2 summary table should be reviewed for other metrics discrepancies when differences between annotation's score are lower than 1000 points.

- **Inner ring**: Represents the area calculated from the radar plot for each annotation represented.

  -The area value is also shown as a numerical value in the legend.
  -This is the annotation quality's **global score** — higher means better. **Check Plot interpretation (step 7) for more information about this global score.**

- **Outer Column sets**: Represents **in-depth annotation metrics.**

  -BUSCO sections: Higher % Complete (Single) means better. If there are no valid biological explanations (e.g., polyploidy), a **high % Complete, Duplicated could indicate assembly problems**, such as redundant contigs generated by pooling DNA from multiple individuals. **High % Fragmented could indicate genome assembly continuity problems**, but this could also be explained by annotation errors if the average CDS length is low.

-**Transposons**: Here is represented the % of transcript models classified as PTeMte (potential TEs). Usually, higher values means worse, but sometimes these sequences are flagged as TEs because they are genes originated from TEs domestication. Check DeTEnGA output file "{species}_TE_summary.csv" list of sequences IDs for further analysis on these sequences.

-**Gene models and transcript models**: Represents how many models are in each annotation. This columns are **normalized by the highest number of models present in all annotations**. For example, if you are representing annotation A with 200 gene models and annotation B with 50 gene models, Annotation B's bar will be 4 times lower than A. Higher means better, but **extreme differences could be explained by annotation problems**, e.g. Helixer doesn't group isoforms into genes, so total number of gene models could be potentially higher than the actual number of organism's genes.

-**Transcripts with both sides UTR**: Represents the % of each annotation's transcripts that have both UTR'5 and UTR'3 annotated. Higher means better.

-**Average CDS length**: Represents the average CDS length. These values are **normalized by the highest Average CDS length**, so if Annotation A is 1000bp and annotation B is 250bp annotation A bar should be 4 times higher. Higher means better, but **extreme differences could be explained by annotation errors**, e.g. incorrect merging of genes as a single gene model.

11 **Plot interpretation**

Different annotation methodologies have their strengths and weakness. Although GAQET plot calculates and overall quality score that could help in selecting a *best* annotation, this score is based in own our selected parameters with an equally weighted score. In future releases we are going to implement a system to set different weights for each scoring parameter, but **it is strongly advised to check each QC parameter on its own before selecting one or more annotations to perform downstream analyses.**

We are going to use **Figure 1 in Step 6** as an example of SGA's QC analysis. First, we have three different methods at the annotation's QC top level:

- **EviANN** (5) (2956 points), that uses evidence based methods like RNA-seq and protein alignments against the genome assembly.
- **Helixer** (6) (24391 points), an *ab initio* annotation method using Deep Neural Networks.

- **Araport 11** (7) (24307 points), the gold standard annotation for *Arabidopsis thaliana* made with a combination of evidence based methods and manual review and curation of gene structures.

Main scores differences between EviANN and the other two annotations are mainly driven by more homology matches against Swissprot Database and by a Higher number of predicted proteins. There is a difference of less than 1000 points between any pair of them, so before selecting and annotation as the best a in-deep review of other features should be performed:

- Total number of gene models predicted is lower in EviANN than the other two annotations.
- Helixer has lower number of transcripts models predicted.
- EviANN has more missing genes under both BUSCO and OMArk analyses.
- EviANN has higher average CDS length than the other two
- Helixer has more UTR in both transcript's flanks annotated than the other two.
- Araport11 have more gene models predicted than EviANN and Helixer

Also, if we inspect the summary metrics we can find that:

- For Helixer and Araport11, the longest gene model is around 27 Kbps, while EviANN's longest model is 411 Kpbs.

In short, **Araport11 would be our choice for most of our downstream analysis**, because it has higher number of gene and transcript models predicted and it's correlated with a higher BUSCO and OMArk completeness, while having the second higher average CDS length.
Some points that are worth to notice:

- Helixer is at the top 3 annotators in this example. The main strong point of this annotation method is that it doesn't need any prior evidence data and is one of the fastest annotators available, although new and promising contenders are starting to arise like ANNEvo.
- GeMoMA (8) is homology-based gene prediction program that works by aligning protein sequences against genome assemblies. This methodology has some problems that are represented in our figure. First, UTRs can't be predicted with

protein alignments. The second one is that it doesn't take into account repetitive masking of genome assemblies, so it tends to annotate transposable elements as genes as show in the figure.

- Stringtie+Transdecoder (9,10) only uses as evidence RNA-seq alignments and it's very dependent on how complete is the representation of the organism's transcriptomic diversity on these datasets. That could explain why BUSCO and OMArk completness results are significantly lower than other methodologies.

Protocol references

1. The UniProt Consortium (2025) UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.*, **53**(D1), D609–D617. **doi: 10.1093/nar/gkae1010**

2. Nevers, Y., Warwick Vesztrocy, A., Rossier, V. *et al.* (2025) Quality assessment of gene repertoire annotations with OMArk. *Nat. Biotechnol.*, **43**, 124–133. **doi:10.1038/s41587-024-02147-w**

3. Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E.M. & Kriventseva, E.V. (2025) OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res.*, **53**(D1), D516–D522. **doi:10.1093/nar/gkae987**

4. Paysan-Lafosse, T., Andreeva, A., Blum, M., Chuguransky, S., Grego, T., Lazaro Pinto, B., Salazar, G.A., Bileschi, M.L., Llinares-López, F., Meng-Papaxanthos, L., Colwell, L.J., Grishin, N.V., Schaeffer, R.D., Clementel, D., Tosatto, S.C.E., Sonnhammer, E., Wood, V. & Bateman, A. (2025) The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res.*, **53**(D1), D523–D534. **doi.org.10.1093/nar/gkae997**

5. Zimin, A.V., Puiu, D., Pertea, M., Yorke, J.A. & Salzberg, S.L. (2025) Efficient evidence-based genome annotation with EviAnn. *bioRxiv* [Preprint]. **doi:10.1101/2025.05.07.652745**

6. Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A.P.M. & Denton, A.K. (2020) Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics*, **36**(22-23), 5291–5298. **doi.10.1093/bioinformatics/btaa1044**

7. Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S. & Town, C.D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.*, **89**(4), 789-804. **doi:10.1111/tpj.13415**

8. Keilwagen, J., Hartung, F. & Grau, J. (2019) GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.*, **1962**, 161-177. **doi:10.1007/978-1-4939-9173-0_9**

9. Pertea, M., Pertea, G., Antonescu, C. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290-295. **doi:10.1038/nbt.3122**

10. Haas, B.J. (n.d.) TransDecoder. Available at: https://github.com/TransDecoder/TransDecoder

11 . Ye, K., Zhang, P., Xu, T., Wang, S., Yang, X., Sun, P., Jia, P., Wang, B., Bush, S. & Ning, Z. (2025) Highly accurate *ab initio* gene annotation with ANNEVO. *Research Square* [Preprint]. **doi:10.21203/rs.3.rs-6402260/v1**