Guidelines for annotating using the ORCAE annotation portal

Getting started

Access via:

- http://bioinformatics.psb.ugent.be/ -> genomes -> Genome of interest -> ORCAE
- http://bioinformatics.psb.ugent.be/orcae/

The website is best viewed with Firefox but can be viewed with any browser (although the layout might differ slightly with other browsers). If you want to use the full functionality of the portal (to view protein alignments and to work on gene structure using Artemis and/or GenomeView) you will need to have Java Web Start installed.

Access to the annotation section is restricted to the annotators so you will need to register and log in before you're able to annotate the genome. Registered users have full access to all sections of the portal. Once you have logged in, you start a new session and you don't have to fill in your login information again as long as your session is active. Note that when you are idle (not doing anything while being logged in) for too long your session will expire and for security reasons you will be logged out automatically, if you want to continue working you will have to login again.

To be invited to create an account you should send a request by email to the project coordinator or to the general email address beg-orcae@psb.ugent.be

After logging in you will be redirected to the home page of your account. There you will find an overview of all the genomes you have access to. For each genome you will be presented with a list of links to the different subsections of the annotation portal (eg. Annotation, Search, Workbench, etc.). Also, a list will be provided of your 5 most recent modifications to that genome.

When you edit a locus a new record will be created in the database. The previous records are not deleted, however, so the database retains all previous annotations. You can always look back at previous versions of a locus using the history. To view an older version, select the date of modification from the dropdown menu. Note, however, that when you modify a record it is always the latest version that is modified (even if you were looking at an older version).

As soon as you modified a locus you become the 'owner' of that locus and your name will be shown in the annotator field. As soon as the default owner name (*Eugene*) is no longer present the monitoring system will be active for that locus. This means that the owner of a locus will automatically be notified by mail if someone has changed that particular locus.

Annotating a gene

This section provides a basic guide to the steps that should be followed to annotate a gene.

Selecting the genes to annotate

Annotation of gene function is based on identifying similarity between a gene in the genome and a gene of known function in the public databases. There are two possible starting points for this. You can use a keyword search (eg. 'transcription factor') to extract genes that have

been annotated with this term during the automatic annotation. This approach may, however, miss genes of interest that had not been annotated with the exact term used in the keyword search. Alternatively, a protein (or DNA) sequence from the public databases can be used to search for related proteins in the genome using Blast. This approach should find all the homologues of a given sequence, provided that there has not been too much sequence divergence. Consider Evalues of e-3 to and e-20 to be a reasonable cut-off, but note that this does not guarantee that the gene is a homologue of the query sequence. The score value given by Blast can also be used to assess similarity. A rule of thumb is that it should be at least 3 times greater than the length of the alignment given by blast (corresponding to about 30% identity over 80 amino acids). Where possible, try to annotate complete gene families or all the enzymes involved in a metabolic pathway. KEGG pathways can be a useful reference source for the latter.

Keyword search. For the keyword search, type the search term into the Keyword Search box on the Search page and click on 'Submit search' (all fields will be searched for the keyword). You will then be provided with the list of genes that have been annotated with the keyword. You can view/annotate ('view') them or transfer them to your workbench ('workbench') to work on them later.

Blast search. For a blast search, you will first need to select a sequence from the public databases for which the function is known or predicted with high confidence (the query sequence). Note that old Genbank entries (with accession numbers of the form X12345) can be particularly interesting for predicting gene function because these old entries were often generated as part of experimental studies. It is also helpful if the query sequence is from a species that is not too phylogenetically distant from the species being annotated. Paste the sequence into the window on the Blast page, select the Blast program (Blastp for a protein) and the local database you want to search, and click on 'Search'. Your blast result will be presented on a new page. From here, you can click on the LocusID (eg. Esi0001_0004) which is highlighted in green to view/annotate the gene(s) retrieved by blast. If you don't get a result from the blast search, this may be because you have used an inappropriate blast program. Note that genes can also be transferred to your workbench from the annotation window by clicking on 'workbench' at the top of the page.

Verifying homology

Additional analyses should be carried out so that you can be more confident that you have identified the functional homologue of a gene. Reciprocal blasts (blasting the *Ectocarpus* gene against Uniprot or Genbank) will provide support for a match if the *Ectocarpus* gene pulls out sequences with the same functional annotation as the initial query gene that was used to search the *Ectocarpus* genome.

The best evidence for functional homology is obtained by building a phylogenic tree using the sequences recovered by Blast and showing that the *Ectocarpus* gene clusters within a group of genes with the same function. Very basic trees can now be generated on the NCBI site.
Once you have identified a gene with the function that you are looking for, you can proceed to the annotation of that gene. This is done on the 'Gene Page' (see below).

Missing genes

If you don't find a match to the gene you are searching for, this may mean that the *Ectocarpus* genome does not contain a homologue. To verify this you should search all the sequence contigs using the Blast page because the gene may be present but not represented among the

gene models generated by Eugene. If no match is found in the database, then it is likely that the gene is absent from the genome (although it is also formally possible that the gene was not sequenced or that it is present but too divergent to be detected by Blast).

Annotating a gene using the Gene Page

There are no mandatory fields to be filled in. This means that you can enter annotations in just one field or in all the annotatable fields and this information will be entered into the database. To add an annotation to a locus click on the 'Modify This Record' button. When you click the 'Modify this Record' button (also available in the side-bar menu) a pop-up window will be opened. This way you can make modifications while still having all the info available from the view page. In the edit section most of the fields of the view page are editable. To submit your changes click on the 'Submit Your Changes' button on the Edit page. After submitting your modifications, they will be validated and the edit page will close automatically after 10 sec (or click 'continue'). On closure of the edit page the view page will reload to show the newly entered information. If you modified the structure of a gene model, that locus will be temporarily locked (status: in update). During this time the similarity information (blast, InterPro, ESTs, …) will be updated, when this is finished the locus will again become available for edit (the context view is updated immediately and will reflect the new situation). Duration of an update is 15-20min depending on server load and protein length.

The gene page is divided into several sections (Annotator, Gene Function, etc). Each section is described in detail below. Note that a comment field is available in each section of the gene page. Here you can enter general comments relevant to that section of the gene page. Use this field to provide additional information about the genes and to note things like conflicts in the annotations provided, errors you are unable to correct, etc. When the conflict is with an annotation entered by another manual annotator, or when you have a question/remark regarding previous manual annotations, you are encouraged to communicate directly with the previous annotator (his/her contact information can be found in the Annotator block).

Graphical overview block

The context overview is the first picture on each gene page. It represents the genomic context of the locus you are currently viewing. Depicted in the overview are the gene models of the selected locus and 2 loci up and downstream of the selected locus. The view is always centred on the selected locus (also indicated by a green box around the gene models). And is automatically rescaled to fit 5 loci (the absolute length of the genomic region shown can differ from locus to locus and is reported below the picture). Genes originated from the forward strand are drawn above the black line (= the genomic sequence), the ones from the reverse strand beneath the line. The black line can be interrupted by red regions; these denote gaps in the sequence. Protein coding genes are coloured blue; RNA genes have a yellow colour.
It's possible to browse the sequence using this view. Clicking on a gene model in the picture will redirect you to that locus page. At the bottom of the figure there are some arrow shaped icons. These are also for browsing: click on the single arrow to move the view one locus in the direction of the arrow. Clicking on the 'double' arrow will shift the view in such a way that the last (or first, depending on the arrow direction) gene model on the current view becomes the first (or last) in the next view (move in steps of 5 loci).

Annotator block

On each gene locus page there is an Annotator section that provides the name and contact information of the person who last modified this locus. The 'default' is *eugene3.4* which indicates that that the annotation presented is that generated by Eugene and that this locus has not been modified by an expert annotator yet. You do not have to enter your annotator information. It will be added automatically as soon as you modify a locus.

Gene Function block

The Gene Function block groups everything related to the function of the gene. These include things like the gene name, KOG and GO identifiers.

- **Short name**
  When the gene acronym is known, it can be entered in this field. The gene acronym is a short code of 3-letters followed by a number or letter that indicates the gene function. Nuclear genes are written in uppercase (eg. UBA1, CDC25, TAP42, STCH). It doesn't matter if you don't feel confident enough to provide a name, in this case the LocusID (eg. Esi0010_0004) will be used to identify the gene.
  A link is provided to go to the gene short name page. This is a top-level page in ORCAE which means it is not linked to a specific genome but accessible from different genome sections. The purpose of this page is to guide people in assigning short names for their genes. Use this page to check if some name already exists or to find the commonly used short name for a gene function.
- **Alternative names**
  If for a particular gene there exists alternative names (either acronyms or full names) they can be entered in this field.
- **Definition**
  This corresponds to the fasta header line and is equivalent to the short description of a gene that you find in Genbank. It should be a short (one-line) text description of the gene that indicates its function. You do not have to add the species name in square brackets nor the locus id, these will be added automatically in the final release.
  Examples: ATP citrate lyase, Zuotin-related chaperone
  Only use words like "putative" if you want to indicate that the functional assignment is uncertain (ie. do not use it systematically to indicate a match with a gene that has this function in the database).
  Note that matches to hypothetical proteins should be annotated "conserved hypothetical protein".
- **Additional functional description**
  Here you can enter a more elaborate description of the gene function. It can be as detailed or as general as you want, as long as the information is accurate and useful to researchers not familiar with this type of protein.
  Example: "PKC is calcium-independent, phospholipid-dependent, serine- and threonine-specific enzyme. PKC is activated by diacylglycerol which in turn phosphorylates a range of cellular proteins. PKC also serves as the receptor for phorbol esters, a class of tumor promoters."
  Alternative names for the gene can also be entered here. It is important to include all alternative names if possible to facilitate future keyword searches.
- **PubMed ID**
  If possible, you should enter Pubmed IDs for the most important publications that report about the specific gene you are working on (ie. in the species being annotated). Alternatively you can also enter Pubmed IDs for publications about orthologous genes from other species that helped you to specify this gene's function. Do not provide full references, the Pubmed identifier (eg. 12024044) is sufficient.

You can enter more than one publication by separating the Pubmed IDs with a semicolon (;).

- **EC number / EC description**
  In this field you can enter/modify the EC number specifying the biochemical activity of the encoded protein. Only the EC number is editable (the description will be added automatically based on the number you provide). See the list of websites at the end of this guide for further information.
  Examples: 1.1.1.8 , 2.4.1.244
- **KOGid / KOGclass / KOG description**
  The Clusters of Orthologous Groups (COG) Database at the NCBI is a collection of sets of orthologous genes from complete genome sequences (including sets of orthologues from eukaryotic genomes, which are termed eukaryotic orthologous groups or KOGs). If you can assign the gene you are working on to an existing KOG, enter the ID in the KOG ID field (again only the ID field should be entered here, all the other information will be added based on the number you provide). See the list of websites at the end of this guide for further information.

The comments box in this section can be used to enter additional information such as relationships to paralogs and orthologs, interaction or subcellular location data, comments about phylogenetic origin, etc.

Gene Ontology block
If available, the GO assignments for each GO-tree separately will be provided in this section. Most of the GO assignments are automatically derived from the InterProScan result (InterPro2Go). Clicking on the GO-id will redirect you to the AmiGO page of that GO-id. See the list of websites at the end of this guide for further information.

Protein Domain block
This section shows all the protein domains that were found in the protein. Domains are mapped using InterProScan, making use of all the public databases. For each domain, the page provides the domain name, the database used to find this domain and a description of the domain.

A schematic representation of the mapped domains is provided at the top of this block. It depicts the relative positions (in aminoacids) of the different domain hits within the protein sequence. Each domain has a different colour (note that the colours do not reflect the score of the hit).

Protein Homologs block
This section shows proteins that share similarity with the protein being annotated. The proteins shown were retrieved by Blast with an Evalue lower than 1e-5 (note that only the best 10 hits are shown if more than ten matches were available with an Evalue of < 1e-5). For each hit the gene name and description as well as the scores are presented. Clicking on the 'View Blast' button will launch a bl2seq of the locus that is being annotated against the specified protein allowing visualisation of the blast alignment.

The schema at the beginning of this block is an overview of a multiple alignment of the query protein and the reported hits. Alignments are constructed with the Muscle program. The gene of interest is colored in blue, the protein hits in green. Boxes represent aligned protein

sequence and the dotted lines are gaps introduced to optimise the alignment. Vertical grey lines indicate the splice site junctions in reference of the protein sequence.
If you want a more detailed view of the alignment you can click on the 'View in JalView' link. This will open a new window showing the aligned protein sequences (JalView). The JalView editor also allows simple tree construction possibilities. Note that you cannot edit this alignment.

Gene Structure block
This section provides the positions of the exons and (when EST sequence is available) the 3' and 5' untranslated regions with respect to the sequence scaffold (supercontig). These positions are given as coordinates. Each exon or untranslated region is indicated by two coordinates separated by two points (eg. 344..387). Exons are separated by commas (eg. 344..387, 566..601). Untranslated regions occur at the beginning or end of the gene and are separated from coding regions by a semicolon (eg. 298..343;344..387 would be the first exon with the start codon being positioned at 344). Next to the coordinates also the sequence type (eg. mRNA, SeCys, ..) and the strand are reported. The quality tag is an indication of the confidence that the structure is the correct one (from 1 to 5).
For a detailed view of the gene structure in either GenomeView or Artemini, click the 'View in GenomeView' respc. 'View in Artemini' link.

In the schema exons are represented by thick, dark blue blocks, untranslated regions by thin, light blue block and introns by arched lines. The strand on which the gene is found is also indicated (+ or -). Note that the schemas are always drawn in the 5' to 3', so genes on the minus strand are reverted. Hovering over the exons/introns will show their length in nucleotides.

The comments box in this section can be used to enter additional structural information such as splice variants, overlaps with neighbouring genes, clustering with genes of related function, etc.

Tiling Array Block (Not available for all genomes!!)
In this graphic a representation of the expression values on the tiling array are provided as a bar-chart. The bars are plotted on the gene structure (note the 5'-3' orientation). A bar is colored green if the expression value is higher than the control values. Otherwise it's colored in red. Clicking the image will redirect you to the data provider homepage.
Graphical representation of the tiling array data mapped onto the gene model.There is a context of 100 bases around the gene model (= tiling array data start 100bases before the model and ends 100bases after it.) The bars denote the expression level. Green color indicates that the experiment value is higher than the control. Otherwise they are colored red. The direction of the gene reflects the strand it is on. They are drawn 5'-3' if located on the forward strand or 3'-5' for the reverse strand.
Click in the grey area to link to the systemix site for a more detailed view (showing the same region) of the expression data.

CDS block
Here you can find the coding sequence (CDS) of the proposed model. You do not have to enter this manually. When you modify the structure of a gene with Artemis, it will be updated automatically.
Clicking on the 'Blast' button will launch a blast against the nr_dna database of NCBI.

Protein block
This shows the translated CDS, ie. the predicted protein sequence. Again, this is generated automatically from the CDS. Clicking on the 'Blast' button will launch a blast of the protein sequence against the Genbank non redundant protein database.

- Signal peptide
  When the protein has a signal peptide present you can enter its sequence in this field. This sequence has to be a part of the protein itself. If no signal peptide is detected you can indicate this in the comments box.

Associated ESTs/cDNAs Block
ESTs and cDNAs that align to the same genomic region as the gene model are provided in this section. If an EST fits the model it is coloured green, if it doesn't fit the model it is coloured red and an explanation is given as to why it doesn't fit the model (or why it was rejected for gene prediction). ESTs that are derived from the same clone (5'-3' couples) are linked with a dashed line in the graphical overview.

The Artemini / GenomeView applet

The gene structure can be altered by using the online Artemis tool (Artemini) or GenomeView in the edit-section. Artemis is a gene visualisation and editing application developed by the Sanger institute. You can obtain some help with using this application from http://www.sanger.ac.uk/Software/Artemis/v7/manual/ . GenomeView is a new genome annotation tool developed by our own group. Opposite to Artemini it can visualise many more types of data. More info can be found at http://genomeview.sourceforge.net/ .
Use the Artemis window to edit the gene structure and save your modifications to enter them into the database. Note that the window only shows part of the supercontig but the coordinates are the same as on the supercontig. When you start an Artemis session the locus will be temporarily unavailable for other annotators to edit. Be sure, when you have finished working in Artemis, to close the Artemis window so that the locus will again be available for other annotators.
It's possible to enter 'new' genes (or to split existing ones). To do so just make a new CDS in artemini and it will be entered in the DB as a new locus. You don't have to enter a locusID for the new gene, this will be done automatically (the new ID will be the last ID on the contig +1). When you split a gene, the one spanning the largest genomic region will inherit the locusID while the shorter one(s) receive new ID(s). You can even create new genes in introns of other genes. Also different sorts of RNA genes can be entered (tRNA, rRNA, miRNA, …), refer to the list provide in the artemini edit view to check which kind of RNA are accepted. (Also RNA genes will receive a locusID as described for the protein coding genes).
To help you in the annotation you now have the possibility to create many more 'features' eg. intron, repeat, … , see the dropdown list in the top left corner of the edit screen for the full list of possibilities. Keep in mind that only the CDS and RNA features will end up in the database, all others will be ignored! Also less common type of proteins can be entered in BOGAS through the artemini/GenomeView as for example SelenoCysteine proteins. To do so create a 'normal' gene but add the qualifier */type='SeCys'* in the gene edit window of artemini. This way some of the basic checks done on submitted gene structures will be skipped (eg. The in-frame stop-check in case of SeCys proteins). Also pseudo genes can be submitted this way. Therefore add the qualifier */pseudo* or */type='pseudo'* in the gene edit window (almost all checks will be skipped). Proteins submitted this way will be tagged accordingly (normal genes have tag 'mRNA', Secys have 'Secys', …).

If you modify gene structures, then loci that overlap with your gene will become obsolete (=permanently locked for edit) and will accordingly be flagged in the DB. Overlap is determined by sequence comparison. If a neighbouring gene has more than 50 bases (with 98% similarity) in common with your gene that whole locus will be considered as obsolete, this also happens if your gene does not cover all the exons of that gene. If you encounter such a case and want to keep the not covered exons in the database, you split of the not covered exons and with the leftover exons you create a new gene (CDS). Genes that have become obsolete will no longer be included in the previous/next navigation nor in the context overview. They are however not deleted but you can only go to their view page by surfing directly to them (eg. by searching for that locus).

The Workbench

As a registered user you will have the possibility to use a workbench. This is used to select and group the loci you are currently working on. You can add several loci to your workbench. When you add a locus to your workbench it will be locked for other annotators (they can't edit it, but they can still view it) until you delete that locus from the workbench.

The Blast Page

You can enter a sequence to blast by pasting it in the window or by uploading a file. After submitting the blast ('Search') you will redirected to the result page. If you searched against the protein or CDS database you can directly go to the Gene page of a hit by clicking on the identifier (eg. Esi0001_0004) or you can recover the sequence of the protein or CDS by clicking on F (eg the F of > F Esi0001_0004). For more detailed information about using this page click on the '?' on top of the page.
It's also possible to link directly from the output of a blast against the genome to locus view pages in ORCAE. To do so click the scaffoldID in the blast result, than a search will be done looking for genes that are covered by the hits in that section of your blast result. If multiple loci are covered then a list of the results will be reported (if no models are present in the blast-hit regions, then nothing is returned).

The Search Page

On the search page you can search for loci based on several different criteria. You can enter a search term in just one of the different sections (eg. KOG or Protein domain) or in more than one section at a time. You can control the way the results from the different sections influence the final result by using the 'AND' 'OR' 'NOT' buttons. You can also use these terms within one field, written in capital! (eg. Esi0011_0023 AND Esi0045_0127). Your search string will be split on the occurrence of those words and each substring will be searched separately in the database. The use of a wildcard ('*') is allowed.
When using the keyword search field, be aware that the use of eg. the AND operators does not imply that the terms have to be found in the same field, one hit can come from the protein homologs information and the other from a gene structure comment. If you want both terms to be present in a single information field you can do: wordA*wordB . When you use the AND in a more specific search field (eg. domain description) than off course the search is restricted to only that field.

You can also search on genomic position. Filling in just a contigID will return all genes on that contig. But by using the region field you can query a specific range on a contig (or all contigs), the range/region is entered as a pair of numbers separated by anything that is not a number. You can combine several regions by the use of the OR operator.
If the result of a search is a single locus you will not see it in a list but you will see directly the locus view page for the result.

The Gene Names Page

The intention of this section is to keep track of the short names for genes. You can browse the list by clicking on the letter links.
You are able to search the existing names (on name or description) and to add new names. To enter new names you will have to provide a short name and a description for that name. It is not possible to change existing names or descriptions. If you feel that some changes need to be done to the list please contact beg-orcae@psb.ugent.be .

The Wiki

You will find a link to the *Ectocarpus* Genome Annotation Wiki on the home page. Use the Wiki to post any information that you think may be relevant to the annotation process. This can include, for example:

General comments about your annotation
Including, for example, the number of genes in a gene family and the structure of the family in terms of sub-groups, how the size of a family compares with other species, functional inferences, etc. Please also make a note of genes that you have searched for but have not found in the genome. This sort of general, genome-scale information is very important.

Your Most Interesting Findings
Note any interesting or strange findings in the Wiki. These can include, for example, a new gene family, an unusual domain structure, evidence for horizontal gene transfer, an unsuspected metabolic pathway, functional gene clustering, shortest exon, alternatively spliced variants with biological significance, overlapping genes, gene in an intron, unusual phylogeny, etc. This information will be extremely valuable when it comes to writing the genome paper.

Notes on the assembly
If you find two parts of a gene on two different contigs this could be useful information to improve the assembly.

Text and figures
At a later stage, the wiki can also be used to post text and figures that can be directly exploited for the genome paper.

Note that the Wiki will include separate pages for each gene family being annotated. Additional pages can be added, either to add additional gene families or for specific annotation questions such as gene duplications, comparative genomics, phylogenomics, etc. Please contact us if you have any suggestions regarding the Wiki. Please report any probable bacterial contamination on the "Report Assembly Problems" page.

Please include your name and the date with all entries to the Wiki to facilitate exchanges between annotators. This is important because there are often multiple annotators for specific gene families. For help entering your data go to the link given on the Wiki home page (http://wiki.splitbrain.org/wiki:dokuwiki)

Useful web pages

Here is a list of useful web pages:

NCBI blastp	http://www.ncbi.nlm.nih.gov.gate1.inist.fr/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on
Blast expasy	http://expasy.org/tools/blast/
Uniprot	http://beta.uniprot.org/
Chlamy gene nomenclature	http://www.chlamy.org/nomenclature.html
EC number	http://www.expasy.ch/enzyme/
KOGnitor	http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html (don't use Firefox)
GO browser	http://www.ebi.ac.uk/ego/
AmiGO	http://amigo.geneontology.org/cgi-bin/amigo/go.cgi?&link=static
KEGG	http://www.genome.jp/kegg/
SignalP	http://www.cbs.dtu.dk/services/SignalP/
PSORT (old version)	http://psort.ims.u-tokyo.ac.jp/form.html
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/
TMHMM	http://www.sbc.su.se/~melen/TMHMMfix/
Interproscan	http://www.ebi.ac.uk/InterProScan/
pfam	http://pfam.sanger.ac.uk/search?tab=searchSequenceBlock
Prosite	http://www.expasy.ch/prosite/
Artemis help	http://www.sanger.ac.uk/Software/Artemis/v7/manual/
Wiki help	http://wiki.splitbrain.org/wiki:dokuwiki