JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

de.NBI

EDGAR - A public database for comparative genomics

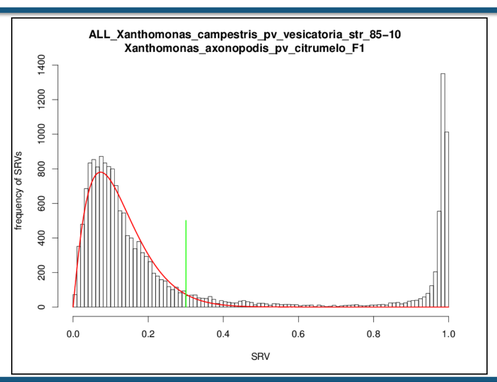
Jochen Blom, Julian Kreis, Sebastian Spänig, Alexander Goesmann
Justus-Liebig-Universität Gießen, Bioinformatics and Systems Biology (Gießen, DE)

The EDGAR platform

ABSTRACT:
The deployment of next generation sequencing approaches has caused a rapid increase in the number of completely sequenced genomes. As one result of this development, it is now feasible to analyze not only single genomes, but large groups of related genomes in a comparative approach. A main task in comparative genomics is the identification of orthologous genes in different genomes and the classification of genes as core genes or singletons.
To support these studies EDGAR - „Efficient Database framework for comparative Genome Analyses using BLAST score Ratios“ - was developed. Using a generic orthology criterion based on the distribution of alignment hits within a genus EDGAR is designed to automatically perform genome comparisons in a high throughput approach. Comparative analyses for 2072 genomes across 161 genera taken from the NCBI genomes database were conducted with the software and the results were integrated into an underlying database. EDGAR provides several analysis and visualization features and significantly simplifies the comparative analysis of related genomes. The web-based user interface offers Venn diagrams, synteny plots, and a comparative view of the genomic neighborhood of orthologous genes.
Recently, the software was extended with various new features like statistical and phylogenetic analyses, replicon grouping options, and second level analyses of meta-genesets. EDGAR calculates phylogenetic trees as well as Average Amino Acid Identity (AAI) matrices based on all genes of the core genome, providing a solid basis for both analyses. Thus the software supports a quick survey of evolutionary relationships and simplifies the process of obtaining new biological insights into the differential gene content of kindred genomes. EDGAR is among the most established tools in the field of comparative genomics. In addition to the publicly available projects more than 200 private projects with more than 5000 analyzed genomes have been computed during the last 5 years.

Orthology estimation

For the computation of a generic orthology criterion EDGAR deploys the so called BLAST Score Ratio Values (SRVs) suggested by Lerat et al. (2003). Instead of using the absolute bit scores provided by the BLAST algorithm, the SRV method uses a normalization approach by relating all bit scores of a protein to the bit score of an alignment of the sequence against itself, as such a BLAST hit gives the maximum bit score possible. So, a SRV is defined as the ratio (Observed score / Maximum score), thus giving a value in the range [0; 1].

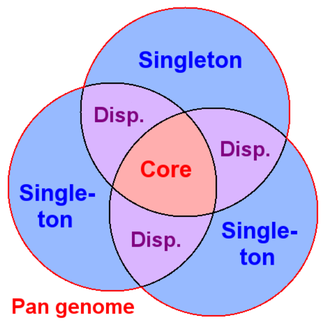


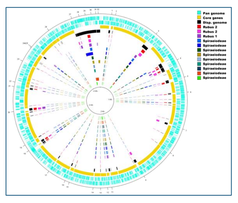
The distribution of bit scores from an all-against-all comparison of the genes of two related genomes shows a bimodal pattern. The first peak represents random matches, while the second peak at higher similarities represents true homologous sequences. To remove all nonspecific hits, a cutoff value has to be identified that defines the boundaries of the first peak and thus the set of random hits that should be removed.
To find a cutoff for the comparison of two genomes, the number of BLAST hits with a given SRV is summed up and represented in a histogram. A beta distribution is calculated from the mean and standard deviation of the observed SRVs < 0.4, and the 97% quantile of this density function is taken as cutoff that defines the end of the first peak and thus the cutoff for the pairwise genome comparison.
Cutoffs are calculated in this way for all possible genome pairs of a comparison set, and the final cutoff for an EDGAR project is estimated via majority decision.

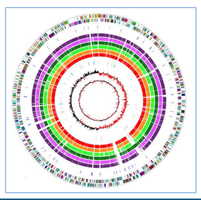
Genomic subset

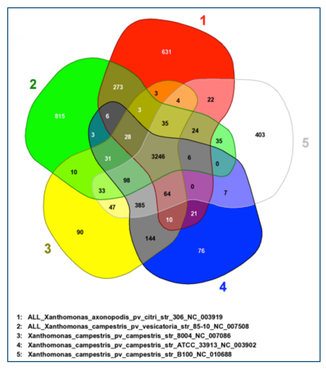
Genomic subsets calculated by EDGAR:

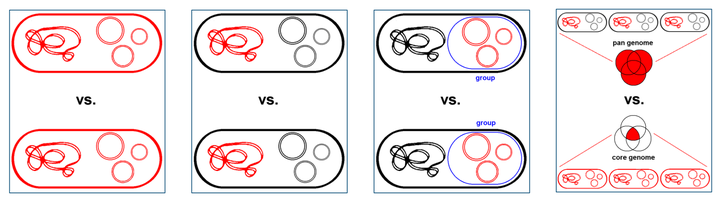
- **Core genome: Conserved among complete set**
- **Singleton genes: Only in one genome**
- **Dispensable genome: Shared in several genomes**
- **Pan genome: Entirety of unique genes of a set**











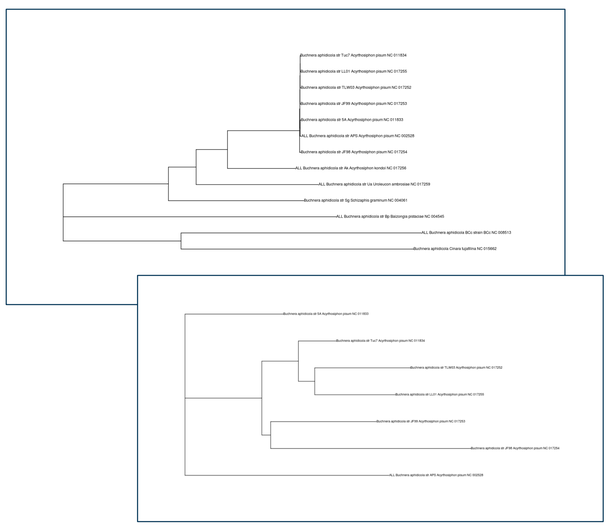
EDGAR allows comparison on different levels. Users can compare complete organisms, single replicons, groups of replicons, or even meta-datasets like previously defined core genomes.

Phylogeny

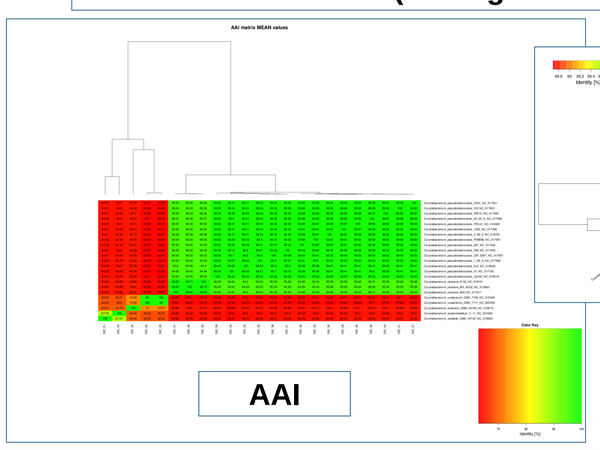
Create phylogenetic trees using following pipeline:

- **Calculate core genome**
- **Multiple alignment of core genes using MUSCLE**
- **Concatenate matching parts, generate huge multiple alignment**
- **Create phylogenetic tree of this alignment using PHYLIP / Neighbor joining**

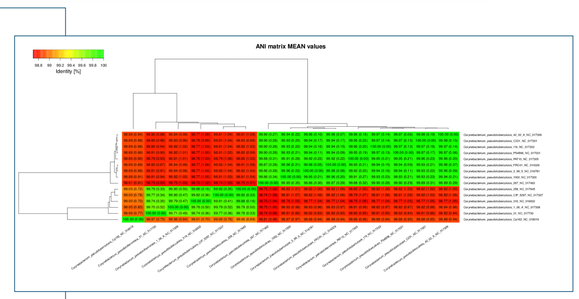
User defined subtrees can be calculated in the web interface.



Further phylogeny features: AAI (Average Amino Acid Identity) and ANI (Average Nucleotide Identity)

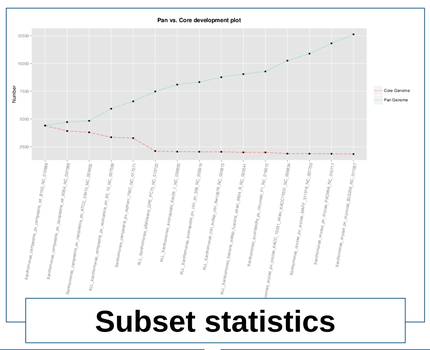


AAI

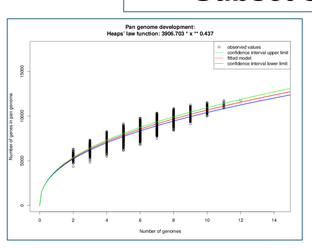


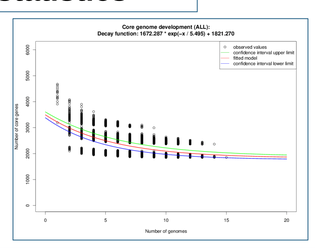
ANI

Further features

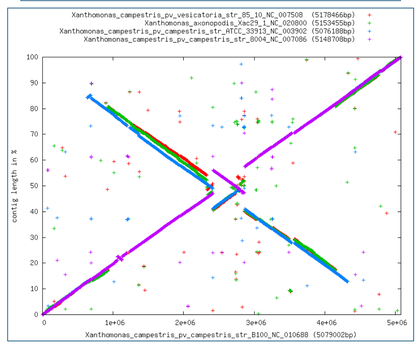


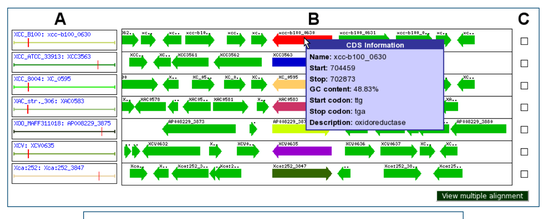
Subset statistics





Synteny plots

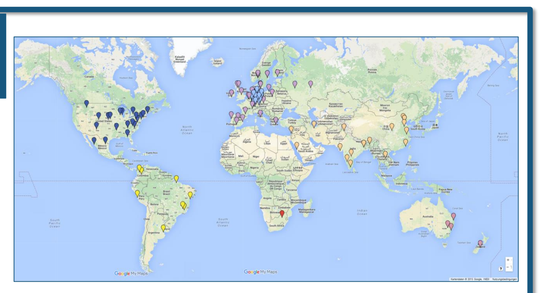




Comparative view

Summary

- **Robust and fast identification of orthologous genes**
- **Classification of genes in genomic subsets (core/pan/singleton)**
- **Public projects for all genera with more than three finished genomes**
- **2072 genomes across 161 genera in public database**
- **>5000 genomes in >200 private projects**



EDGAR users worldwide

EDGAR Web server:

http://edgar.computational.bio

Contact:
Jochen Blom
jochen.blom@computational.bio
http://edgar.computational.bio

Projektträger Jülich
Forschungszentrum Jülich

Bundesministerium für Bildung und Forschung

http://www.denbi.de
http://computational.bio