W22–W28 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 20 April 2016*
doi: 10.1093/nar/gkw255

EDGAR 2.0: an enhanced software platform for comparative gene content analyses

Jochen Blom[1,*], Julian Kreis[1], Sebastian Spänig[1], Tobias Juhre[1], Claire Bertelli[2,3], Corinna Ernst[4] and Alexander Goesmann[1]

[1]Bioinformatics & Systems Biology, Justus-Liebig-University Giessen, 35392 Giessen, Hesse, Germany, [2]Institute of Microbiology, University Hospital Center and University of Lausanne, 1011 Lausanne, VD, Switzerland, [3]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, VD, Switzerland and [4]Center for Familial Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, University of Cologne, 50931 Cologne, NRW, Germany

Received January 31, 2016; Revised March 22, 2016; Accepted April 2, 2016

ABSTRACT

The rapidly increasing availability of microbial genome sequences has led to a growing demand for bioinformatics software tools that support the functional analysis based on the comparison of closely related genomes. By utilizing comparative approaches on gene level it is possible to gain insights into the core genes which represent the set of shared features for a set of organisms under study. *Vice versa* singleton genes can be identified to elucidate the specific properties of an individual genome. Since initial publication, the EDGAR platform has become one of the most established software tools in the field of comparative genomics. Over the last years, the software has been continuously improved and a large number of new analysis features have been added. For the new version, EDGAR 2.0, the gene orthology estimation approach was newly designed and completely re-implemented. Among other new features, EDGAR 2.0 provides extended phylogenetic analysis features like AAI (Average Amino Acid Identity) and ANI (Average Nucleotide Identity) matrices, genome set size statistics and modernized visualizations like interactive synteny plots or Venn diagrams. Thereby, the software supports a quick and user-friendly survey of evolutionary relationships between microbial genomes and simplifies the process of obtaining new biological insights into their differential gene content. All features are offered to the scientific community via a web-based and therefore platform-independent user interface, which allows easy browsing of precomputed datasets. The web server is accessible at http://edgar.computational.bio.

INTRODUCTION

The revolutionary improvements in high-throughput DNA sequencing during the last 10 years have dramatically increased the availability of complete and draft microbial genome sequences. As a result, thousands of sequences are now available in the public sequence repositories, and tens of thousands of sequencing projects are ongoing. Thanks to this treasure of available data, the comparative analysis of the differential gene content of genomes quickly became a routine task in modern genomics. Especially the estimation of the core genome, the pan genome and singleton genes as defined by Tettelin *et al.* (1) and Medini *et al.* (2) are important steps in the analysis of groups of genomes. Several software platforms for comparative gene content analyses have been developed in the last decade like IMG (3), MicrobesOnline (4), MBGD (5) or OrthologeDB (6). IMG and MicrobesOnline are designed as general purpose genomics databases for a broad variety of genomic information, but provide only a limited range of comparative analysis features. MBGD and OrthologeDB are focused on comparative genomics, but both don not place much emphasis on result visualization and don't provide phylogenetic analyses. To support comparative gene content analyses combined with visual result representation, the software EDGAR (7) was developed. The initial version of EDGAR, referred to as 'EDGAR 1.0' in the following, supported only a limited range of analysis features, namely the calculation of genomic subsets and visualizations like Venn diagrams and pairwise synteny plots. The collection of features provided by EDGAR has been extended to inlcude a range of sophisticated analyses since then, with a focus on phylogenetic and statistical analyses. Existing features have been modernized and updated continuously. In the following chapters the updated and new features will be presented in detail.

*To whom correspondence should be addressed. Tel: +49 641 99 35803; Fax: +49 641 99 35809; Email: jochen.blom@computational.bio.uni-giessen.de

TECHNICAL UPGRADES IN EDGAR 2.0

Since the publication of EDGAR 1.0 in 2009, several changes of the back-end and front-end of the software have been realized.

In EDGAR 1.0, all mathematical calculations were implemented in Perl. Most graphics were created using gnuplot (http://gnuplot.info) and Perl/CGI graphics. For the release of EDGAR 2.0, the visualization frameworks and libraries were changed to allow more up-to-date interactive graphics. All statistical and curve fitting calculations are now implemented in the statistical computing language *R* ((8), https://www.r-project.org/), as well as the respective plots. For interactive result visualization, a combination of HTML5, JavaScript in general and the Highcharts (http://www.highcharts.com/) charting library in particular was used. The database back-end was changed from one local SQLite (http://www.sqlite.org/) database per EDGAR project to a central MySQL server (http://www.mysql.com) running the InnoDB storage engine. Project calculations are distributed to a 1000 CPU core compute cluster.

IMPROVED AND MODERNIZED FEATURES

For the high-throughput computation of comparative analyses it is crucial to rely on a robust orthology criterion consistent within the analyzed genome set. For this purpose, EDGAR utilizes the so called BLAST Score Ratio Values (SRVs) suggested by Lerat *et al.* (9). The basic principle used in EDGAR is still the same as described in (7), but significant improvements have been made to the method. A detailed description of the updated orthology calculation of EDGAR 2.0 is provided in the Supplementary Data.

Genomic subset calculation

The main feature of EDGAR was and still is the fast calculation of the genomic subsets defined in the introduction: the core genome, pan genome and singleton genes. All calculations require the selection of one reference genome, and a set of genomes to which the reference should be compared. The reference genome acts as starting point for iterative extension or reduction of the result gene set, which is presented in tabular form. The result table shows the locus tags as well as descriptions of the genes. In addition, result tables now provide multiple alignments of the ortholog sets on nucleotide as well as on protein level. Results can be saved as multiple FASTA file (DNA or protein sequence) or as a TAB separated flat file.

Venn diagrams

Venn diagrams show the number of genes for all possible logical combinations of a selection of genomes. They allow an easy visual inspection of the core genome size and the gene numbers in every subset of the dispensable genome. The EDGAR web interface features the creation of Venn diagrams with an upper limit of five genomes because the number of logical combinations within a Venn diagram of higher order results in too many areas for a meaningful graphical representation. Genome comparisons of a higher order are possible, though, via a new interface that enables calculation of any possible intersection of any arbitrary number of genomes. In this interface the user can select single genomes as included, excluded, or ignored, and EDGAR will calculate the gene set matching the query and present the results in tabular form. The diagram layout has been notably improved since EDGAR 1.0, providing more even sized areas and an improved coloring scheme. An example of the new Venn diagram layout used in EDGAR 2.0 is shown in the Supplementary Data.

Synteny plots

Synteny describes the co-localization of genes on a stretch of DNA. A synteny plot showing the conservation of gene order among several genomes is an easy way to identify large scale evolutionary events like genome rearrangements. The original EDGAR web server provided an interface to create synteny plots of pairs of genomes based on the stop positions of genes that were identified as being orthologous. Plots were generated as static images with gnuplot. In EDGAR 2.0 synteny plots can be created for up to 20 genomes at a time. The genomes are compared to a selected reference genome, and a track is plotted in a different color for each of them (see Figure 1). The individual tracks can be switched on and off, and the order in which the genome tracks are superimposed on each other can be changed dynamically. Thus, the synteny plot is now a highly interactive tool for the analysis of large scale genome rearrangements.

Genome browser

To gain more convenient visual access to the genomic neighborhood of orthologous genes, a new genome browser was added to the EDGAR web interface as replacement for the comparative viewer presented in the original publication. In EDGAR 2.0, we introduce a JavaScript and HTML5 based Genome Browser. This interactive tool allocates the same color to orthologous genes, and shows the genomic context in a window of 20 kb. Thereby the genome browser allows rapid detection of the presence or absence of orthologous genes and variations in the gene order. Additionally, users can interactively realign the genes in the genome browser window by clicking on a gene. Moreover, a multiple alignment of a selected gene set can be generated, allowing biologists to verify the ortholog relationship. All gene sets visible in the 20 kb window at a given time are additionally presented in tabular form below the interactive genome browser.

NEW FEATURES ADDED TO THE EDGAR WEB SERVER

Besides the presented improvements, EDGAR 2.0 also provides novel features and concepts that have not been available before. For example, in EDGAR 1.0 only chromosomes could be compared, but organisms with multiple replicons could not be handled properly. In EDGAR 2.0, multi-replicon-organisms are fully supported, and all analysis features can be run either on the single replicons, or an a virtual container comprising all genes of

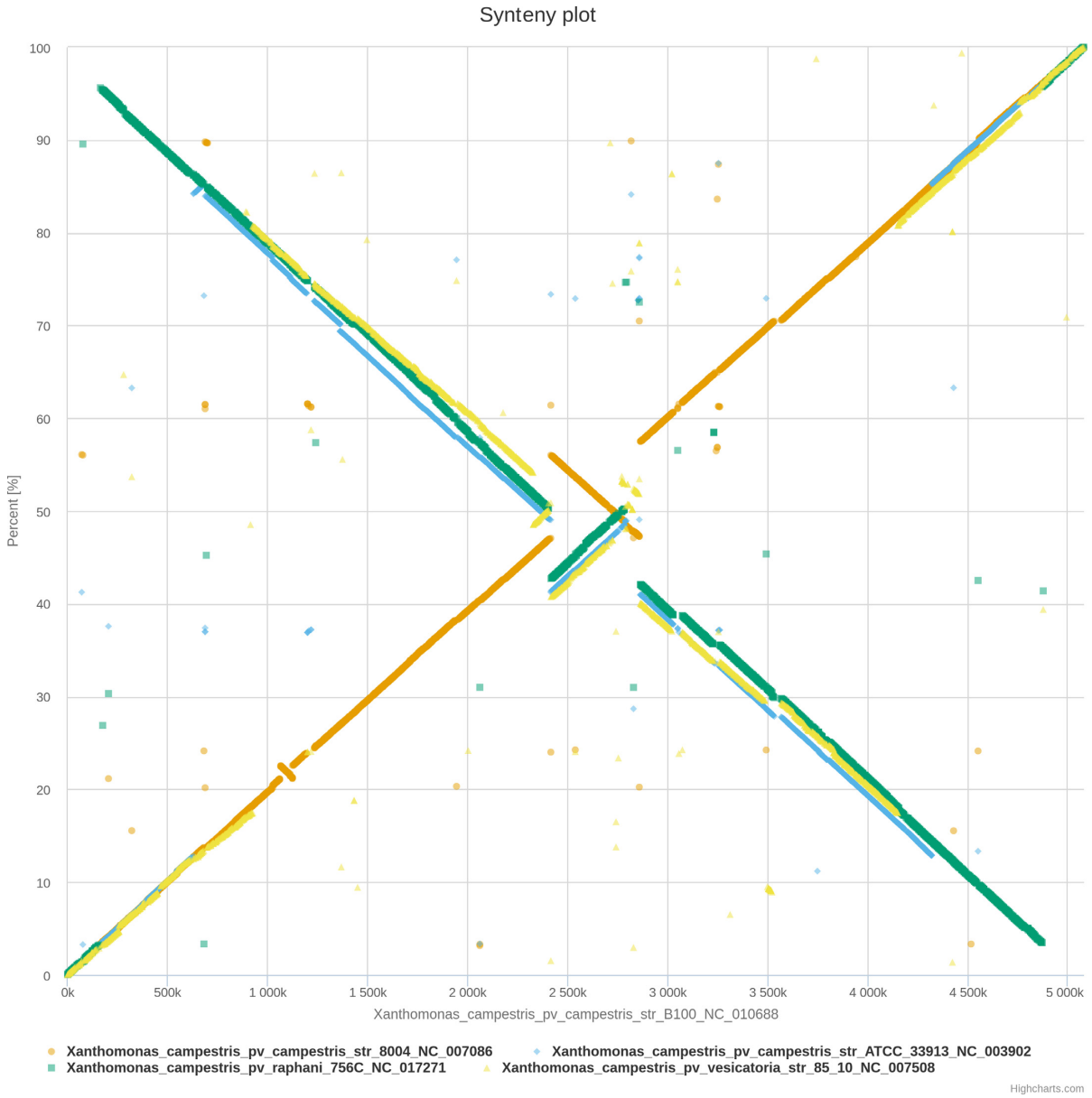

Figure 1. Synteny plot of four *Xanthomonas campestris* chromosomes compared to *X. campestris pv. campestris* strain B100.

an organism. These containers are automatically generated during the EDGAR project calculation and are named "ALL_<organism name >". In the following the most important new features of EDGAR 2.0 are presented in detail.

Genomic subset statistics

A calculated genomic subset, e.g. a core genome calculated on a specific set of genomes, is always only a snapshot of the situation for the given genome set. One possible solution to obtain a more comprehensive estimation of genomic subset sizes is to calculate the respective numbers for every possible combination of all available genomes and to use the resulting data to extrapolate how subset sizes would develop for an infinite number of genomes. Mathematical approaches for genomic subset extrapolation were proposed by Tettelin *et al.* in 2005 (1) and 2008 (10) and are now implemented in EDGAR 2.0.

Core genome and singleton development extrapolation. The development of the core genome size for increasing numbers of genomes can be predicted by a curve fitting approach using an exponential decay function. An identical approach is used to extrapolate the development of the expected number of singletons, thus, to facilitate the mathematical description only the core genome development calculation is described here.

If k genomes are available, one estimates the number of core genes for all $\binom{n}{k}$ possible permutations of the genomes. Subsequently, the number of core genes is plotted as a function of the number of compared genomes. Using a non-linear least squares curve fitting approach, an exponential

decay function of the form:

$$f(n) = c \cdot \exp\left(\frac{-n}{\tau}\right) + \Omega \quad (1)$$

is fitted to the data, where c is the amplitude of the exponential function, n is the number of compared genomes, τ is the decay constant that defines the speed at which f converges to its asymptotic value and Ω is the extrapolated size of the core genome for $n \to \infty$. Thus, the Ω value indicates how well the core genome size of the currently available genomes represents the 'real' core genome size of the analyzed genus. Figure 2A shows the core genome development plot for 14 *Xanthomonas* genomes.

Pan genome development extrapolation. The development of the pan genome size can be estimated using a Heaps' law function. Heaps' law is an empirical law mainly used in linguistics describing the number of distinct words in a document (or a set of documents) as a function of the document length. When an increasing number of texts is analyzed, the number of different words grows according to a sub-linear power law of the total number of scanned words. The development of the pan genome shows a comparable development and can be extrapolated by a power law of the form:

$$f(n) = c \cdot n^{\gamma}, \quad (2)$$

where n is the number of compared genomes, c is a proportionality constant and γ the growth exponent. As in the core genome and singleton statistics the parameters c and γ can be estimated by non-linear least squares curve fitting to the data points from a calculation of the pan genome size for all possible permutations of the available genomes. An exemplary pan genome size extrapolation for 14 *Xanthomonas* genomes is shown in Figure 2B.

Pan vs. Core development plot. When the aforementioned statistical features are used, it is crucial to ensure that consistent genomic data are used. The results can be strongly influenced by genomes with a high evolutionary distance to the rest of the dataset. Furthermore, the calculations can be disturbed by genomes with highly differing gene content due to poor gene prediction accuracy or highly fragmented draft genomes. To identify such outliers, the pan versus core development plot is the ideal tool. Starting with one genome, a sequence of core and pan genome sizes is calculated by iteratively adding one genome at a time to the comparison in a user-defined order. Outliers can be easily detected in the resulting pan versus core plot as demonstrated by Figure 3.

Phylogenetic analysis features

While phylogenetic analyses were not part of the web server in EDGAR 1.0, a phylogenetic tree of all available genomes is now calculated by default for all EDGAR projects. For that purpose, EDGAR 2.0 uses the phylogenetic analysis pipeline developed on the basis of the ideas of Zdobnov *et al.* (11) which was described in the use case in (7). This pipeline analyzes the phylogenetic relationships between genomes based on the thousands of orthologous genes in the complete core genome. Multiple alignments of each orthologous gene set of the core genome are calculated using the MUSCLE software (12). The resulting alignments are concatenated to one large complete core alignment which is used to create a phylogenetic tree using the neighbor joining method as implemented in the PHYLIP package (13).

Subtrees. For some genera subbranches in the phylogenetic tree might be hard to resolve due to the close phylogenetic proximity of a certain species, e.g. for *Mycobacterium tuberculosis* within the *Mycobacterium* genus. For such cases EDGAR 2.0 offers an interface to calculate phylogenetic trees of a subset of genomes in the project. This feature enables a more detailed view of the selected subset of genomes. At the same time the reliability of the result is increased compared to the parent tree, since the size of the core genome, which is the basis of the tree calculation, increases for the reduced genome set.

ANI and AAI. While the computation of a phylogenetic tree based on the complete core genome shows good results, it is still a computationally intensive task. Two different approaches toward a phylogenetic evaluation based on the increasing availability of whole-genome sequences were proposed by Konstantinidis *et al.* (14–16), i.e. the average amino acid identity (AAI) and the average nucleotide identity (ANI). Both methods are provided by the EDGAR 2.0 web server.

For the AAI method, the average AAIs of all conserved genes in the core genome as computed by the BLAST algorithm (17) are collected. The results can be easily extracted from the EDGAR database. ANI values are computed as described in (18) and as implemented in the popular JSpecies package (19). For both methods, the resulting phylogenetic distance values are arranged in an AAI/ANI matrix, clustered according to their distance patterns and visualized as heatmaps. The heatmap images as well as the raw AAI/ANI values can be exported from the web server.

Retrieval of orthologous gene sets

The EDGAR 2.0 web server provides several ways to search and retrieve data. One of them is the retrieval of orthologous gene sets, which allows users to define a set of query genes, e.g. all genes of an operon. All genes that are orthologous to the query genes in all selected comparison genomes are identified and presented as detailed tables. This feature is thus the perfect tool to quickly find genes of interest for scientists focusing on a certain type of genes.

Upstream motif search. The EDGAR 2.0 database not only stores all coding sequences of a set of genomes, but also stores up to 400 bp of the sequence upstream of the gene start. This allows a search for conserved motifs in these upstream regions like, the Pribnow box (20), σ^B-binding motifs (21), cold shock protein binding motifs (22), etc.

Inspired by the GECO software (23), an upstream motif search was implemented in EDGAR 2.0 using the *fuzznuc* software provided by the EMBOSS package (24). Users can search for PROSITE-style nucleotide patterns, either in an exact search or with up to two allowed mismatches. Genes

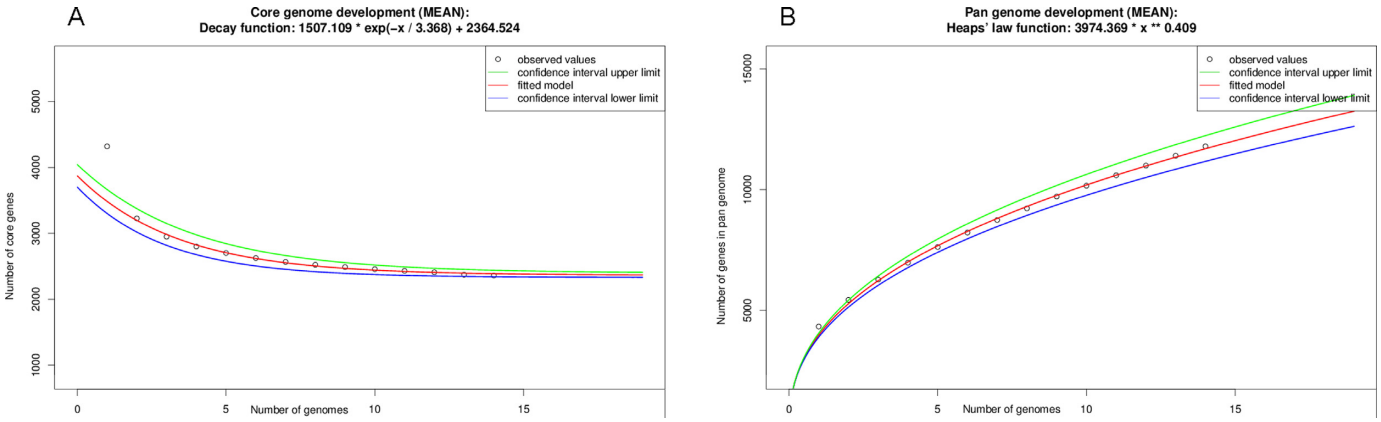

Figure 2. (**A**) Core genome development plot for 14 *Xanthomonas* genomes. The red curve shows the fitted exponential decay function, blue and green curves indicate the upper and lower boundary of the 95% confidence interval. The extrapolated core genome size is 2364 genes. (**B**) Pan genome development plot for 14 *Xanthomonas* genomes. The red curve shows the fitted exponential Heaps' law function, blue and green curves indicate the upper and lower boundary of the 95% confidence interval. Based on these results the pan genome is considered to be open with a growth exponent of 0.409.

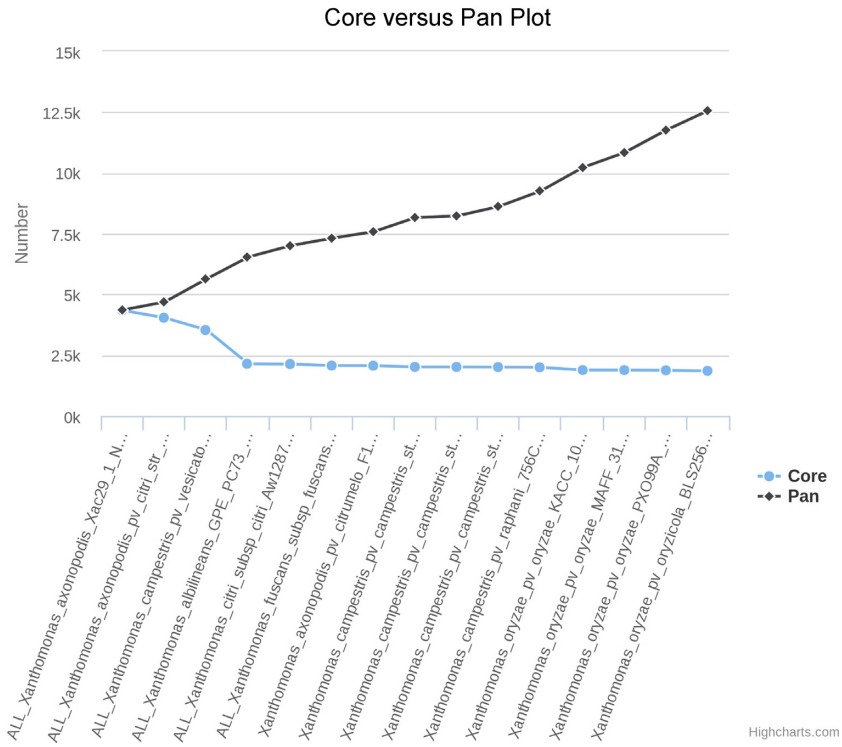

Figure 3. Pan versus core development plot of 15 *Xanthomonas campestris* genomes. The drastic drop of the core genome size with the introduction of *Xanthomonas albilineans* strain GPE PC73 is clearly visible. The outlier status of this genome is confirmed by the phylogenetic tree.

that have the query motif in their upstream region will be displayed in a table showing the exact position of the motif.

Core HMM scan. As EDGAR stores huge amounts of data from millions of BLAST comparisons, the question arose how this data could be used to analyze data that was not included in the EDGAR project calculation. One approach in this direction is the creation of profile Hidden Markov Models (HMMs, (25)) from orthologous gene sets. During the calculation of an eDGAR project, the protein sequences of all sets of orthologous genes with more than five members are aligned using MUSCLE (12) and a profile HMM is created using the HMMER3 package (26). The resulting HMM database can be queried in the EDGAR 2.0 web interface.

Higher level analysis features

As already described, EDGAR provides data structures to compare single replicons or complete organisms. For comparing the gene content of all plasmids of one organism to the genes of all plasmids of another organism, a higher level of abstraction is needed. For such cases EDGAR al-

lows users to create groups consisting of all genes of these replicons.

These groups are well suited to join replicons from one organism, but if replicons from different organisms need to be grouped, the orthologous genes from the involved organisms act as artificial paralogs and prevent a reasonable analysis. If such multi-organism-groups are needed, e.g if a researcher wishes to compare the gene content of a set of pathogenic bacteria against a set of non-pathogenic bacteria, disjunct gene sets representing a group of organisms are required. A straight forward solution is to calculate the core or pan genome of a set of genomes and to store one representative of each gene set for subsequent calculations. Such non-redundant representations of genomic subsets are called 'meta contigs' and can be created in the EDGAR 2.0. These meta contigs allow higher level comparisons. For instance, it is possible to check which genes of a genome are unique in comparison to the complete pan genome of a genome set.

REQUIREMENTS

The precomputed EDGAR databases can be accessed via the EDGAR 2.0 web server at http://edgar.computational.bio. EDGAR projects are organized on genus level, and an alphabetically ordered overview of available projects is provided on the start page. The EDGAR website is free and open to all users and there is no login requirement.

In addition, for the analysis of unpublished data, password protected private databases can be created on request. For such private databases, any arbitrary collection of annotated genomes can be used. The accepted input files are all DDBJ/ENA/GenBank feature table formats.

Incomplete genomes

EDGAR 2.0 is capable of processing incomplete genomes by joining the contigs of such draft genomes to pseudochromosomes. As the EDGAR method is based on the comparison of coding sequences, incomplete genomes do not pose any technical problems. Nevertheless, one should be aware that every draft genome adds bias to the EDGAR results, as every gap in a sequence may split, truncate or completely mask a gene. Thus, the usage of heavily fragmented genomes or too many draft genomes should be avoided when using the EDGAR platform. This is also the reason why the public EDGAR databases only use completed genomes.

DISCUSSION AND CONCLUDING REMARKS

Since the publication of EDGAR 1.0, the EDGAR web server has become one of the most popular resources for comparative genomics. The number of publicly available projects has increased from 75 genus-based projects with 582 genomes to 167 genera and 2160 genomes. Furthermore, more than 300 private projects are currently provided to users from more than 100 universities and institutes all over the world. The largest EDGAR user base is located in Europe, but more than 50% of the researchers using EDGAR are from outside of Europe. The popularity of the EDGAR service is also reflected by the fact that in 2015 alone more than 10 000 genomes have been processed.

During the last years, the web interface has been modernized with up-to-date graphical visualizations, and the feature set was significantly extended. With the genomic subset statistics, higher level comparison features and several data search and retrieval methods EDGAR 2.0 offers a unique and comprehensive set of comparative analyses which in most cases were developed based on user feedback. The added value of the new features has been proven in numerous studies that successfully used EDGAR for phylogenetic and taxonomic analyses (27). Furthermore EDGAR was used in studies with medical (28), ecological (29,30) or agricultural (31) background.

The EDGAR platform has been continuously developed and improved, and the next upcoming features are already planned. One new concept will be the usage of EDGAR data for genome annotation. Another field for improvement are the phylogenetic analysis features, where more *in silico* genome-to-genome comparison features will be implemented. The main task for the mid-term development of EDGAR will be to replace the data back end once again. While SQlite was sufficient in the 454 sequencing era and the MySQL server works fine for the amounts of data that have to be analyzed today, the ever increasing amounts of data provided by modern sequencing systems make a further stage of development necessary. Thus, a change of the EDGAR data model and the development of a NoSQL data back end have already been started.

With the presented features, EDGAR 2.0 supports a quick survey of evolutionary relationships among microbial organisms, simplifies the search for genes of interest and provides new biological insights into the differential gene content of kindred genomes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors further wish to thank the Bioinformatics Core Facility (BCF) for expert technical support.

FUNDING

The EDGAR platform is financially supported by the German Federal Ministry of Education and Research within the de.NBI network [FKZ 031A533]. Funding for open access charge: Open Access publication fund of the Justus-Liebig-University.

Conflict of interest statement. None declared.

REFERENCES

1. Tettelin,H., Masignani,V., Cieslewicz,M., Donati,C., Medini,D., Ward,N., Angiuoli,S., Crabtree,J., Jones,A., Durkin,A. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13950–13955.
2. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.

3. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Jacob,B., Huang,J., Williams,P. *et al.* (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.*, **40**, D115–D122.
4. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
5. Uchiyama,I., Mihara,M., Nishide,H. and Chiba,H. (2013) MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res.*, **41**, D631–D635.
6. Whiteside,M.D., Winsor,G.L., Laird,M.R. and Brinkman,F.S. (2013) OrthologueDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. *Nucleic Acids Res.*, **41**, D366–D376.
7. Blom,J., Albaum,S.P., Doppmeier,D., Pühler,A., Vorhölter,F.-J., Zakrzewski,M. and Goesmann,A. (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, **10**, 154.
8. R Development Core Team. (2008) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, ISBN 3-900051-07-0.
9. Lerat,E., Daubin,V. and Moran,N. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol.*, **1**, 101–109.
10. Tettelin,H., Riley,D., Cattuto,C. and Medini,D. (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, **11**, 472–477.
11. Zdobnov,E. and Bork,P. (2007) Quantification of insect genome divergence. *Trends Genet.*, **23**, 16–20.
12. Edgar,R. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
13. Felsenstein,J. (1989) PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, **5**, 163–166.
14. Konstantinidis,K.T. and Tiedje,J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.*, **187**, 6258–6264.
15. Konstantinidis,K.T. and Tiedje,J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2567–2572.
16. Konstantinidis,K.T., Ramette,A. and Tiedje,J.M. (2006) The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. Lond B Biol. Sci.*, **361**, 1929–1940.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Goris,J., Konstantinidis,K.T., Klappenbach,J.A., Coenye,T., Vandamme,P. and Tiedje,J.M. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, **57**, 81–91.
19. Richter,M. and Rosselló-Móra,R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 19126–19131.
20. Pribnow,D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 784–788.
21. Hain,T., Hossain,H., Chatterjee,S.S., Machata,S., Volk,U., Wagner,S., Brors,B., Haas,S., Kuenne,C.T., Billion,A. *et al.* (2008) Temporal transcriptomic analysis of the Listeria monocytogenes EGD-e regulon. *BMC Microbiol.*, **8**, 1–12.
22. Morgan,H.P., Estibeiro,P., Wear,M.A., Max,K.E., Heinemann,U., Cubeddu,L., Gallagher,M.P., Sadler,P.J. and Walkinshaw,M.D. (2007) Sequence specificity of single-stranded DNA-binding proteins: a novel DNA microarray approach. *Nucleic Acids Res.*, **35**, e75.
23. Kuenne,C., Ghai,R., Chakraborty,T. and Hain,T. (2007) GECO–linear visualization for comparative genomics. *Bioinformatics*, **23**, 125–126.
24. Rice,P., Longden,I., Bleasby,A. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
25. Eddy,S.R. (1996) Hidden markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
26. Eddy,S. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
27. Borriss,R., Chen,X.-H., Rueckert,C., Blom,J., Becker,A., Baumgarth,B., Fan,B., Pukall,R., Schumann,P., Spröer,C. *et al.* (2011) Relationship of Bacillus amyloliquefaciens clades associated with strains DSM 7T and FZB42T: a proposal for Bacillus amyloliquefaciens subsp. amyloliquefaciens subsp. nov. and Bacillus amyloliquefaciens subsp. plantarum subsp. nov. based on complete genome sequence comparisons. *Int. J. Syst. Evol. Microbiol.*, **61**, 1786–1801.
28. Sangal,V., Blom,J., Sutcliffe,I.C., von Hunolstein,C., Burkovski,A. and Hoskisson,P.A. (2015) Adherence and invasive properties of Corynebacterium diphtheriae strains correlates with the predicted membrane-associated and secreted proteome. *BMC Genomics*, **16**, 1–15.
29. Glaeser,S.P., Imani,J., Alabid,I., Guo,H., Kumar,N., Kämpfer,P., Hardt,M., Blom,J., Goesmann,A., Rothballer,M. *et al.* (2015) Non-pathogenic Rhizobium radiobacter F4 deploys plant beneficial activity independent of its host Piriformospora indica. *ISME J.*, **10**, 871–884.
30. Ngugi,D.K., Blom,J., Stepanauskas,R. and Stingl,U. (2015) Diversification and niche adaptations of Nitrospina-like bacteria in the polyextreme interfaces of Red Sea brines. *ISME J.*, doi:10.1038/ismej.2015.214.
31. Mann,R., Blom,J., Bühlmann,A., Plummer,K., Beer,S., Luck,J., Goesmann,A., Frey,J., Rodoni,B., Duffy,B. *et al.* (2012) Comparative analysis of the Hrp pathogenicity island of Rubus- and Spiraeoideae-infecting Erwinia amylovora strains identifies the IT region as a remnant of an integrative conjugative element. *Gene*, **504**, 6–12.

Published online 01 July 2016 *Nucleic Acids Research, 2016, Vol. WEB SERVER ISSUE 2016, No. 00* **1–2**
doi:10.1093/nar/gkn000

EDGAR 2.0 – Supplemental material.

Jochen Blom [1*], Julian Kreis [1], Sebastian Spänig [1], Tobias Juhre [1], Claire Bertelli [2,3], Corinna Ernst [4], and Alexander Goesmann [1]

[1]Bioinformatics & Systems Biology, Justus-Liebig-University Giessen, Heinrich-Buff-Ring 58, 35392 Giessen, Hesse, Germany and [2]Institute of Microbiology, University Hospital Center and University of Lausanne, 1011 Lausanne, VD, Switzerland and [3]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, VD, Switzerland and [4]Center for Familial Breast and Ovarian Cancer, Medical Faculty, University Hospital Cologne, University of Cologne, 50931 Cologne, NRW, Germany

Received January 31, 2016; Revised –.–.—-; Accepted –.–.—

IMPROVED ORTHOLOGY ESTIMATION

For orthology estimation EDGAR uses bidirectional best BLAST (1) hits (BBHs) with a generic orthology threshold calculated from the similarity statistics of the compared genomes. EDGAR still uses the BLAST Score Ratio Value (SRVs) approach suggested by Lerat *et al.* (2) and described in the initial publication of the EDGAR framework (3). In short, the bit scores of all alignment results provided by the BLAST algorithm are normalized in relation to the maximal achievable bit score, the score of the self-hit of a gene. In contrast to EDGAR 1.0, the threshold estimation from the SRV statistics was changed from a sliding window approach to a statistical approach based on the beta distribution. The number of BLAST hits with a given SRV is still summed up and represented in a histogram for all SRV values, and a beta distribution is calculated from the mean and standard deviation of the observed SRVs within an interval $[0,0.4]$. A 97% quantile of the density function of the beta distribution is defined as the border of the left low SRV score peak and thus as the orthology cutoff for a pairwise genome comparison. The 97% cutoff is based on manual inspection of hundreds of SRV histograms. A typical bimodal SRV histogram with fitted beta distribution is shown in Figure 1.

This procedure is repeated for all possible combinations of compared genomes, resulting in n^2 combinations for a set of n genomes. The final orthology threshold for the complete genome set is generated by a majority decision among this n^2 pairwise cutoffs. An EDGAR project is realized by an all-against-all comparison of all genes of a set of genomes using BLASTP. The resulting BLAST hits are filtered according to the calculated orthology threshold and stored in a MySQL database. The database serves as backend for the EDGAR web server, where all subsequent comparisons are calculated on the fly based on the precomputed BLAST results. EDGAR considers two genes to be orthologous to each other if (A) they have reciprocal best BLAST hits (BBHs), and (B) the SRV values of both single BBHs is above the cutoff. To avoid ambiguities due to identical paralogs within one genome, multiple 100% identical instances of a gene are reduced to one representative during the EDGAR project calculation. The information about filtered genes is tored and will be displayed in EDGAR results.

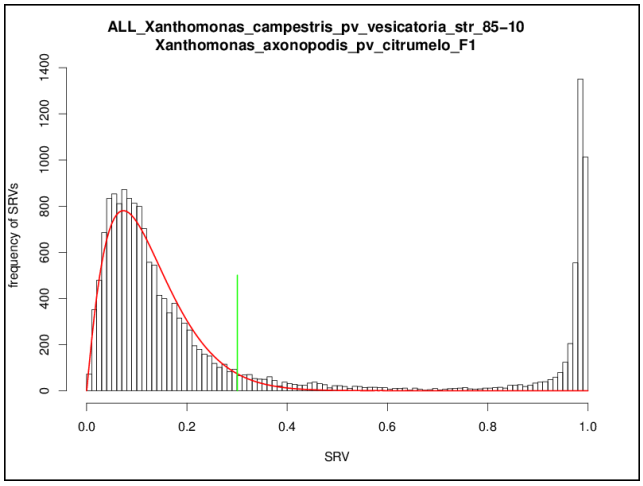

Figure 1. Histogram of a bimodal SRVs distribution resulting from the comparison of two Xanthomonas genomes. The red curve shows a beta distribution calculated for all values between $[0,0.4]$, and the green line shows the 97% quantile of this distribution which is used as cutoff value.

The orthology cutoff generated by this approach is quite strict, as all low quality BLAST hits are filtered out. In an example project with 42 genomes from the genus *Erwinia*, the calculated SRV cutoff is 31. The BLAST comparisons were run with an evalue cutoff of $1e^{-5}$ and generated about 40 million BLAST results. Only ∼7.3 million or 18.25% of these results passed the SRV filter. The mean percent identity of all hits was 73.5% (median 79.0%), and the mean evalue $6.6e^{-9}$ (median $6.0e^{-103}$). This confirms the strictness of the filter. In this way it supports the desired high specificity of the orthology estimation.

*To whom correspondence should be addressed. Tel: +49 641 9935803; Fax: +49 641 9935809; Email: jochen.blom@computational.bio.uni-giessen.de

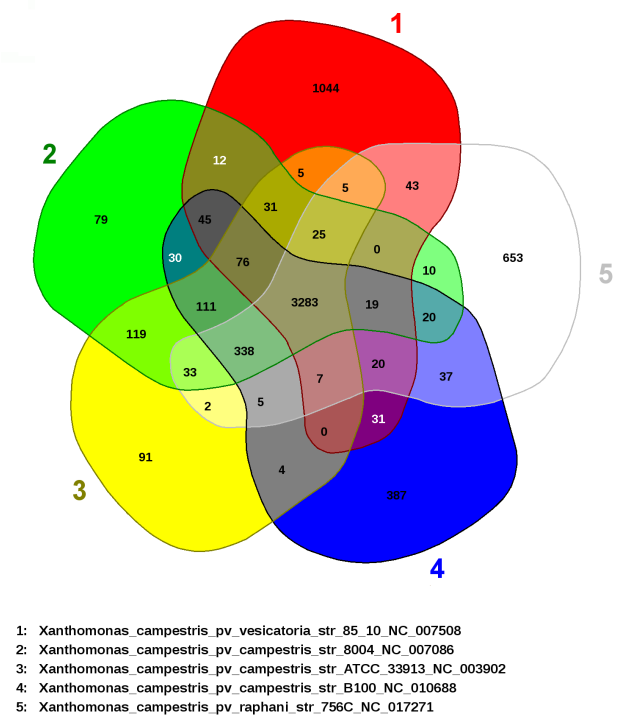

Figure 2. Venn diagram showing the differential gene content of five *Xanthomonas campestris* genomes.

VENN DIAGRAMS

Figure 2 shows the improved Venn diagram layout in EDGAR 2.0. Each included genome has one basic color, and all areas in the Venn diagram representing combinations of genomes are colored in the respective combination color of the included genomes. The numbers within the Venn diagram are now links to a detailed table of all genes that contribute to the respective genome subset. This allows for a detailed inspection of all subsets of the dispensable genome.

REFERENCES

1. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
2. E. Lerat, V. Daubin, and N.A. Moran. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biology*, 1(1):E19, 2003.
3. Jochen Blom, Stefan P Albaum, Daniel Doppmeier, Alfred Pühler, Frank-Jörg Vorhölter, Martha Zakrzewski, and Alexander Goesmann. Edgar: a software framework for the comparative analysis of prokaryotic genomes. *BMC bioinformatics*, 10(1):1, 2009.