`cblaster`: a remote search tool for rapid identification and visualisation of homologous gene clusters

Cameron L.M. Gilchrist[1*], Thomas J. Booth[1], Bram van Wersch[2],
Liana van Grieken[2], Marnix H. Medema[2*], Yit-Heng Chooi[1*]

[1]School of Molecular Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley, 6009

[2]Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands

*Correspondence: cameron.gilchrist@research.uwa.edu.au, marnix.medema@wur.nl, yitheng.chooi@uwa.edu.au

Abstract

Genes involved in coordinated biological pathways, including metabolism, drug resistance and virulence, are often collocalised as gene clusters. Identifying homologous gene clusters aids in the study of their function and evolution, however existing tools are limited to searching local sequence databases. Tools for remotely searching public databases are necessary to keep pace with the rapid growth of online genomic data. Here, we present `cblaster`, a Python based tool to rapidly detect collocated genes in local and remote databases. `cblaster` is easy to use, offering both a command line and a user-friendly graphical user interface (GUI). It generates outputs that enable intuitive visualisations of large datasets, and can be readily incorporated into larger bioinformatic pipelines. `cblaster` is a significant update to the comparative genomics toolbox. `cblaster` source code and documentation is freely available from GitHub under the MIT license (github.com/gamcil/cblaster).

Introduction

Complex biological processes are coordinated through the action of multiple distinct, yet functionally associated genes, which are often found physically collocated within genomic neighbourhoods as gene clusters. Gene clusters have been extensively studied in microbes for their ability to encode traits such as the biosynthesis of secondary metabolites, virulence, drug resistance and xenobiotic degradation; however, gene clustering is a phenomenon observed across all kingdoms of life (Chevrette et al. 2020; Wang et al. 2019; Nützmann et al. 2018; Foflonker and Blaby-Haas 2021). Given the relationship between genetic collocation and cofunctionality (Lee 2003; Michalak 2008), it follows that we can detect gene clusters by directly searching for conserved instances of collocation across taxa. Indeed, tools such as MultiGeneBlast (Medema et al. 2013) and clusterTools (Lorenzo de los Santos and Challis 2019) have been developed precisely for this purpose. These tools take specific queries (nucleotide/protein sequences, Hidden Markov Model (HMM) profiles) and search them against genomic datasets, identifying any

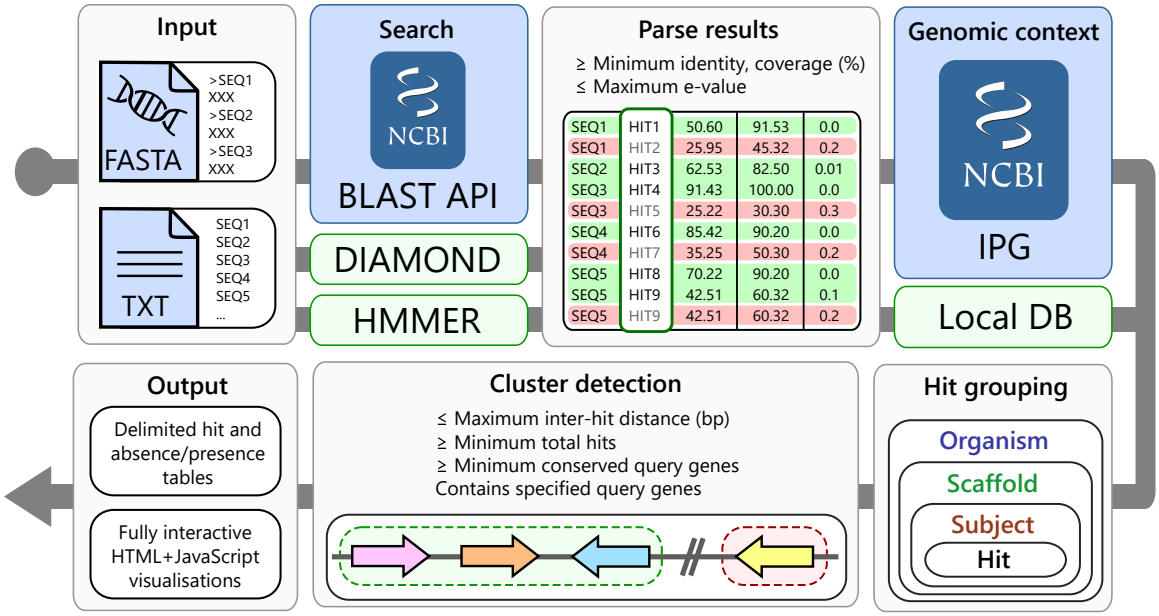


Figure 1: The `cblaster` search workflow. Input sequences are given either as a FASTA file or as a text file containing NCBI sequence accessions. They are then searched against the NCBI's BLAST API or a local DIAMOND database, in remote (blue background) and local (green background) modes, respectively. BLAST hits are filtered according to user defined quality thresholds. Genomic coordinates for each hit are retrieved from the Identical Protein Groups (IPG) resource. Hits are grouped by their corresponding organism, scaffold and subjects. Finally, hit clusters are detected in each scaffold and results are summarised in output tables and visualisations.

instances of collocated query hits. However, their utility is limited by the requirement to build and maintain local databases. In many cases, this is redundant given that most biological sequence data is deposited in publicly accessible online databases like the National Center for Bioinformatics Information (NCBI). It is also increasingly demanding as the amount of available data has grown rapidly in the past decade. Between January 2010 and September 2020, the total number of RefSeq genomes grew from 10,171 to 104,969 (NCBI RefSeq Growth Statistics page). This has two additional consequences: i) maintaining an up-to-date local database has become significantly more difficult; and ii) the size of search outputs has increased significantly. Therefore there is a need for search tools that can leverage remote databases and visualise search outputs in an intuitive manner.

Here, we present `cblaster`, a Python-based tool to rapidly search for collocated protein coding regions remotely by leveraging NCBI APIs, or locally within user-generated databases. `cblaster` is easy to use with both command line and graphical user interfaces (GUI). It generates fully interactive visualisations implemented in JavaScript and HTML that allow the user to intuit patterns from complex datasets, as well as textual outputs that can easily be incorporated into bioinformatic pipelines. `cblaster` is a significant update to the comparative genomic toolbox and is a launch pad for further functional and evolutionary analyses.

Fully remote searches against NCBI sequence databases

The `cblaster` search workflow is detailed in Figure 1 and can be launched either through the command line interface (Table 1) or the GUI (Figure 2). A search begins with the user providing either protein sequences

Table 1: Overview of modules provided by the `cblaster` command-line interface.

Module	Description
`search`	The main `cblaster` search workflow
`makedb`	Create local search databases from personal sequence data
`gne`	Gene neighbourhood estimation of `cblaster` search results
`extract`	Extract sequences from `cblaster` search results
`gui`	Launcher for the `cblaster` graphical user interface (GUI)
`extract_clusters`	Extract detected gene clusters from `cblaster` sessions
`plot_clusters`	Plot detected gene clusters using `clinker` (Gilchrist and Chooi 2021)

(FASTA, GenBank or EMBL format), or a collection of valid NCBI protein sequence identifiers (i.e. accessions, GI numbers), that they believe may form a conserved gene cluster. If the latter is provided, sequences corresponding to each identifier are downloaded using the NCBI Entrez API (NCBI Resource Coordinators 2017) prior to starting the search. Input sequences would typically be taken from the output of a cluster discovery pipeline such as antiSMASH (Blin et al. 2019), or a resource such as the Minimum Information about a Biosynthetic Gene cluster database (MIBiG; Kautsar et al. 2019). Sequences are uploaded to the NCBI BLAST API to launch a new search using BLASTp from the NCBI BLAST+ suite (Camacho et al. 2009; NCBI Resource Coordinators 2017). Optionally, an Entrez search query can be provided to pre-filter the search database, for example to specify a taxonomic group of interest, resulting in vastly reduced search times. Every BLAST search is assigned a unique request identifier (RID) which remains active for 36 hours. A `cblaster` search can be resumed at any point given a valid RID and the corresponding query sequences, such that search results can be retrieved at a later time if desired. This also allows `cblaster` to analyse searches launched through the BLAST website, which can be convenient when expecting long search times due to, for example, many query sequences or unfiltered search databases. `cblaster` can also save search sessions, which can be freely loaded back into the program. The session file contains all data generated by `cblaster` during a search, and can be used to re-detect clusters under new parameters or generate new visualisations without having to completely repeat entire searches.

The BLAST API is repeatedly polled until the search has completed, at which point search results are downloaded and filtered based on user-defined hit quality thresholds (minimum identity, coverage and maximum e-value). These thresholds are especially useful for narrowing down searches when expecting large result datasets. To retrieve the genomic context of remote BLAST hits, `cblaster` leverages the Identical Protein Groups (IPG) resource via the Entrez API (NCBI Resource Coordinators 2017). The IPG database stores links from proteins to their exact genomic coordinates, meaning BLAST hits can be traced back to their genomic origins. To ensure no potential clusters are missed, for example due to annotation error or fragmented genome assembly, all rows in the IPG table are saved and linked back to their original hits. Hits are grouped by their subject proteins, which are in turn grouped by genomic scaffold and organism. Finally, subjects are grouped into clusters if they contain hits to any required query sequences specified by the user (if any), and satisfy thresholds for maximum intergenic distance, minimum size and minimum hits per unique query sequence. After a gene cluster is finalised, it is assigned a unique numeric ID which is reported in the search output, which can be referenced in other `cblaster` modules.

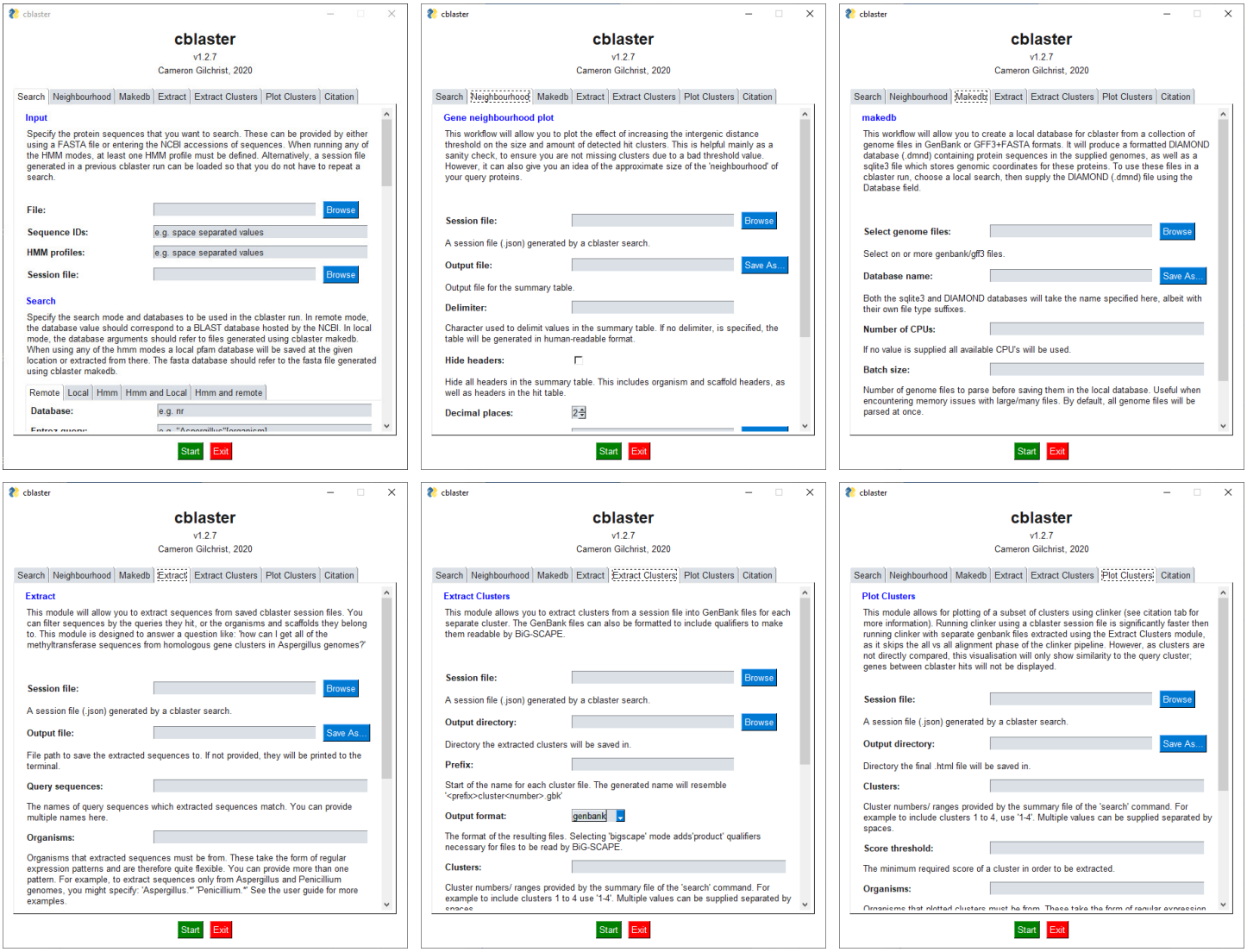


Figure 2: The `cblaster` graphical user interface (GUI). Each panel reprisents a single `cblaster` module: `cblaster search` (top left) for perfoming searches against remote and local databases; `cblaster gne` (top right) for performing genomic neighbourhood estimation; `cblaster makedb` (bottom left) for building local databases from GenBank files and; `cblaster extract` (bottom right) for extracting FASTA files of specific groups of homologues.

Clusters are ranked using the scoring formula previously implemented in MultiGeneBlast (Medema et al. 2013). Briefly, cluster similarity is calculated by

$$S = h + i \cdot s \tag{1}$$

where h is the number of query sequences with BLAST hits, s is the number of contiguous gene pairs with conserved synteny, and i is a weighting factor (default value 0.5) determining the weight of synteny in the similarity score.

Local searches against custom sequence databases

`cblaster` can also search local sequence databases. The local search workflow mirrors that of remote searches, with two key differences: i) the DIAMOND search tool (Buchfink et al. 2015) is used instead of BLASTp; and ii) a local genome database is used in place of the IPG database to facilitate retrieval of the genomic coordinates of hit proteins.

A local search thus requires two databases to be created by the user: a formatted DIAMOND search database and a local genome database. `cblaster` provides a module, `makedb`, which can rapidly generate both databases from a collection of genome sequences simultaneously in a single command. Briefly, genome files are parsed using the BioPython (Cock et al. 2009) or gffutils (https://github.com/daler/gffutils) libraries. `cblaster` then builds a SQLite3 database containing coordinates of genes within each genome, which can be loaded during local searches. Protein translations are extracted to a FASTA file from which the DIAMOND database is built; sequence headers correspond to unique database indices, allowing `cblaster` to retrieve information about sequences hit in a search.

Searching for functional composition using domain profiles

Although it is useful to identify clusters through sequence similarity, in some cases this can be too restrictive. Another approach is to search for profile hidden Markov models (HMMs), which correspond to functional domains. In this way, one can search for gene clusters containing hypothetical functionality as opposed to sequence similarity. This is particularly useful when using a 'retro-biosynthesis' approach, where biosynthetic gene clusters are identified based on functions informed by chemical structures (Cacho et al. 2015). It can also be useful in cases where mutual sequence similarity between members of a protein family is low. For example, microbial terpene synthases typically show only weak sequence similarity, but can be readily identified through profile HMM searches (Komatsu et al. 2008).

Similar functionality has been implemented previously in ClusterTools (Lorenzo de los Santos and Challis 2019), where users can search local sequence databases for combinations of profile HMMs. `cblaster` provides a search mode, `hmm`, where users can perform searches using profile HMMs from the Pfam database (Mistry et al. 2021) as queries instead of protein sequences. To do this, `cblaster` wraps the the HMMER software package (Mistry et al. 2013). In this workflow, query domain profiles are first extracted from the Pfam HMM

profile database using `hmmfetch`, and then searched against a local sequence database (FASTA file of amino acid sequences generated using the `makedb` module) using `hmmsearch`. This requires a local copy of the Pfam database; `cblaster` will automatically download the latest Pfam release if no copy is found. Due to the dependency on the HMMER package, this functionality is currently only supported on Linux and Mac systems. `cblaster` can also perform hybrid searches, where profile HMMs are searched alongside protein sequences.

Estimation of genomic neighbourhood size

Gene cluster structure can vary greatly between organisms. For example, fungal and bacterial biosynthetic gene clusters tend to be tightly packed, though in plants and some exceptions in fungi, genes encoding biosynthetic pathways are loosely clustered (Kessler et al. 2020; Liu et al. 2020a). In other cases, genes may be split across multiple loci as mini subclusters (Bradshaw et al. 2013), or intertwined with clusters encoding other pathways to form superclusters (Wiemann et al. 2013). During cluster detection, `cblaster` uses a user-defined threshold to determine the maximum distance between any two BLAST hits in a cluster. By default, a cluster is finalised if no new hit is found within 20Kbp of the previous hit. However, this value is arbitrary and may be inappropriate for some datasets, and so it is advisable to test the effect of changing this parameter on each dataset being analysed. `cblaster` provides the `gne` module, which leverages its ability to rapidly reload and recompute search sessions to robustly automate this analysis. When provided with a session file, `gne` iteratively performs cluster detection over a user-determined range of intergenic distance threshold values, then plots the total number of predicted clusters, as well as the mean and median cluster size (bp), at each value (Figure 3b; Additional File 2). These plots typically resemble logarithmic growth, steeply rising at low values but gradually levelling off at higher values. This makes it easy to determine sensible cutoffs that avoid liminal regions where small variability would have a large effect on the output.

Comprehensive output and fully interactive visualisations

`cblaster` offers rich summaries and visualisations of search results. `cblaster`'s outputs are available in both human-readable and character-delimited formats, and can be easily incorporated into higher level bioinformatic pipelines. This includes a summary of all detected clusters and a presence/absence table (here termed binary table), which shows the total number of hits in detected clusters per query sequence. By default, clusters are grouped by the organisms and scaffolds they are in; they can optionally be printed in rank order, with the most similar clusters appearing first.

Detected gene clusters can be extracted into GenBank format files directly from `cblaster` search sessions using the `extract_clusters` module. Users can filter clusters by their numeric IDs, the organism and scaffold they appear on, or by a minimum cluster similarity score. Optionally, features within each GenBank file can be formatted for interoperability with the biosynthetic gene similarity clustering and prospecting engine (BiG-SCAPE; Navarro-Muñoz et al. 2020). In a similar vein, sequences within detected clusters can be extracted using the `extract` module. Sequences can be filtered by the query sequence/s they were hit by during the BLAST

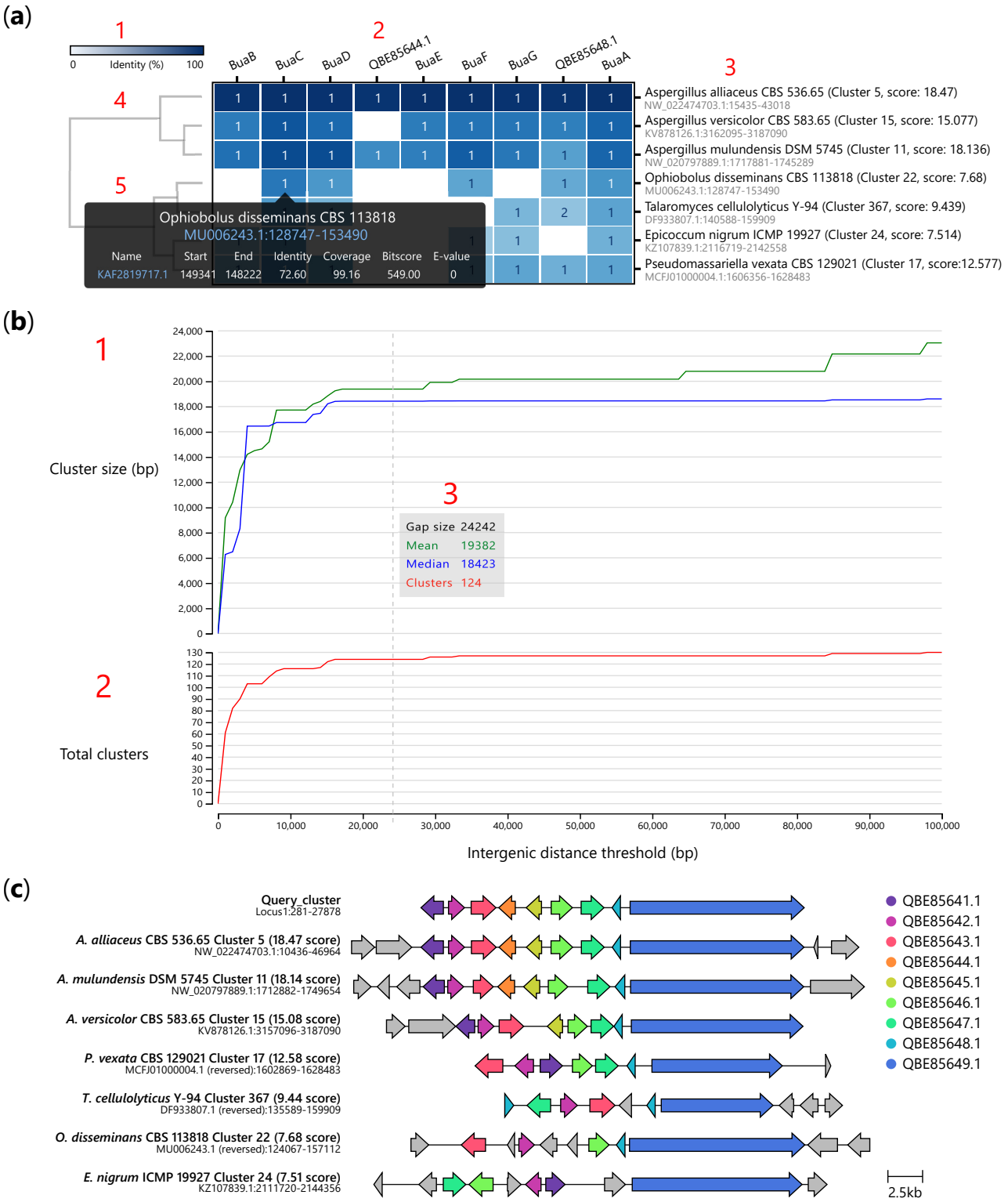


Figure 3: Interactive visualisations generated by `cblaster`. a) Cluster heatmap visualisation of `cblaster` search results: 1) heatmap colour bar indicating 0% (white) to 100% (blue) identity; 2) names of query sequences; 3) names of organism and scaffold locations of hit clusters; 4) dendrogram of hit clusters generated from their identity to query sequences; and 5) cell hover tooltip with detailed hit information including hyperlinks to genomic position on NCBI. b) Gene neighbourhood estimation (GNE) visualisation: 1) Plot of mean and median hit cluster sizes (bp) at different gap sizes; 2) Plot of total clusters at different gap sizes; 3) Hover tooltip showing values of mean and median cluster size (bp) and total clusters at a given gap size. c) Visualisation of gene clusters, inclusive of intermediate genes (grey colour), identified in a) generated using the `clinker` tool via the `plot_clusters` module in `cblaster`.

search, their organism or scaffold. The `extract` module can produce delimited summary tables as well as FASTA format files of all sequences matching the specified filters.

Search results are visualised as a cluster heatmap (Figure 3a; Additional File 1). Clusters are hierarchically clustered based on best hit identity values using the SciPy library (Virtanen et al. 2020). Cells in the heatmap are shaded based on identity. The text inside each cell indicates if query sequences have multiple hits within a cluster. Figures can be freely panned and zoomed; columns (query sequences) and rows (clusters) can be hidden by clicking on their respective labels. Additionally, mousing over a cell in the cluster heatmap generates a tooltip that displays a summary of hits in the cluster for the corresponding query sequence. The tooltip also links to the exact cluster location on the NCBI's graphic genome viewer.

Similarly, `cblaster gne` results are visualised as line charts (Figure 3b) which can be panned and zoomed, with a tooltip showing the number of detected clusters, as well as the mean and median cluster size (bp), at different intergenic distance values.

`cblaster` visualisations are implemented as HTML documents that can be opened in the web browser. These documents contain scalar vector graphics (SVG) images generated using the D3 visualisation library (Bostock et al. 2011), which can be exported for modification in vector image manipulation software. Any changes made within the visualisation are reflected in the exported image. Additionally, `cblaster` can produce fully portable HTML documents can be generated, enabling results to be shared between different computers.

Finally, `cblaster` provides a module for analysing subsets of search results using the `clinker` tool (Gilchrist and Chooi 2021), allowing visual comparison of the structure (inclusive of intermediate genes) of specific clusters of interest. This module mirrors the functionality of the `extract_clusters` module, whereby clusters can be filtered by ID, organism, scaffold and score.

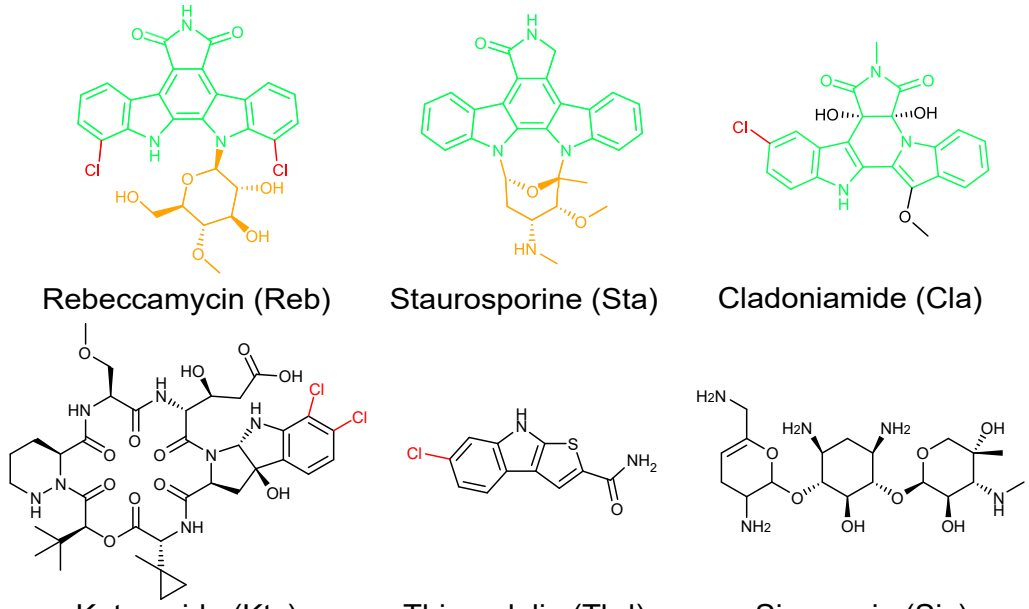
Case studies

Case Study 1: Analysing evolutionary relationships of chromopyrrolic acid derived natural products

To assess `cblaster`'s ability to visualise evolutionary relationships, we searched for homologues of the *reb* BGC against a local database of characterised BGCs. The *reb* BGC is responsible for the production of the indolocarbazole rebeccamycin in strains of the Actinobacterium *Lechevalieria aerocolonigenes* (Sánchez et al. 2002). Indolocarbazoles are biosynthesised through the successive modification of chromopyrrolic acid, a dimer of of tryptophan. Furthermore, they are closely related to other groups of tryptophan derived natural products, most significantly the indolotryptolines. As such, they represent a good case study to demonstrate the ability of `cblaster` to identify and group related families of BGCs. Using the `cblaster makedb` module, a local database was generated from GenBank sequences downloaded from the October 2019 release of the MIBiG database (Kautsar et al. 2019). The database was queried using protein sequences extracted from the published rebbeccamycin BGC (GenBank accession AJ414559) (Sánchez et al. 2002).

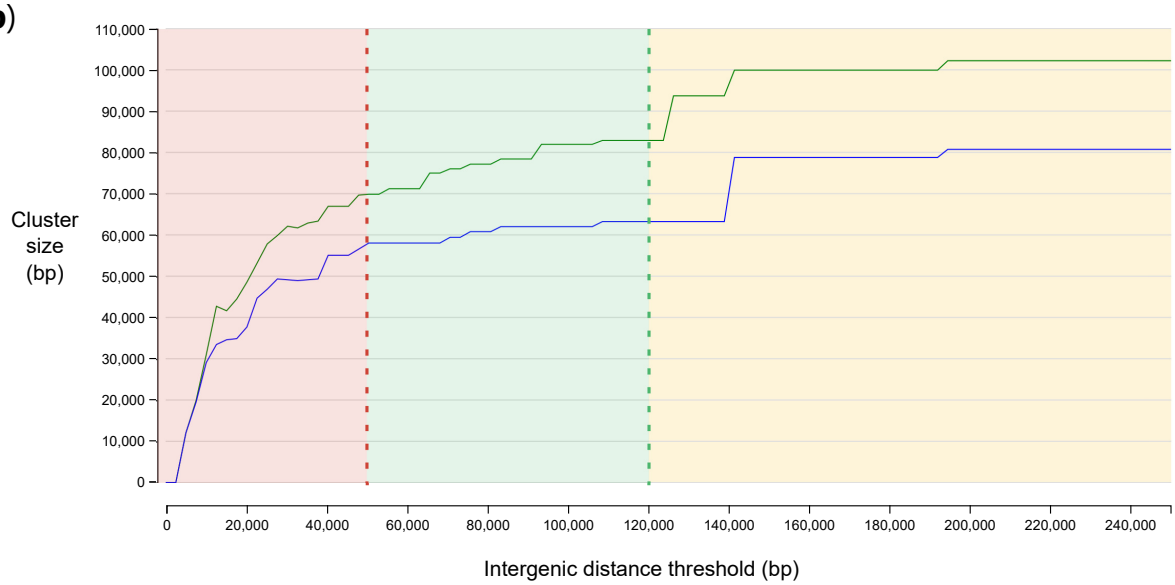
The topology of the resulting dendrogram accurately reflects the previously described groups, indicating

(a)



Reb	T	H	U	F	R	M	P	C	D	O	G
Cla											
Sta											
Ktz											
Thd											
Sis											

(b)



(c)

Figure 4: Caption next page

Figure 4: Application of `cblaster` to case studies in bacteria, plants and fungi. a) The evolutionary relationships between rebeccamycin and other natural products as predicted by `cblaster`. Structural features are highlighted according to conserved biosynthetic proteins (halogenases in red, chromopyrrolic acid biosynthesis in green, glycosyltransferase and methyltransferase in yellow). b) Genome neighbourhood estimation of plant triterpene BGCs. 'Liminal' region of plot shown in red, 'stable' region in green and upper limits in yellow. The accompanying total clusters plot (as in Figure 3b) is ommitted from this figure, but follows the same pattern. c) Using `cblaster` to piece together burnettramic acids BGC in fragmented *A. burnettii* genome.

a distinction between the indolotryptolines, the rebeccamycins and other indolocarbazoles (Figure 4a, Figure S1). Interestingly, `cblaster` identified two additional groups containing *reb* homologues. Firstly, a group of BGCs encoding structurally unrelated compounds sharing tryptophan halogenases (RebH) and flavin reductases (RebF), including kutzneride (Fujimori et al. 2007), thienodolin (Wang et al. 2016) and ulleungmycin (Son et al. 2017). Although structurally diverse, all members of this group contain a halogenated tryptophan motif in the final structure. Secondly, a group sharing homologues of one or both or the *reb* transport proteins (RebT and RebU), including the glycosides gentamicin (Unwin et al. 2004; Huang et al. 2015) and sisomicin (Hong et al. 2009).

Although these are only surface level observations, the results demonstrate the ability of `cblaster` to rapidly highlight both apparent and less conspicuous evolutionary relationships.

Case Study 2: Genomic neighbourhoods and the biosynthesis of triterpenes in plants

Recent studies have highlighted the importance of genomic neighbourhoods in the evolution of plant development and metabolism (Nützmann et al. 2016; Mihelčić et al. 2019). Brassicacaea for example, have evolved to biosynthesise a range of triterpene products through the process of gene shuffling in plastic regions of the genome (Liu et al. 2020c). During triterpene biosynthesis 2,3-oxidosqualine is cyclised by an oxidosqualine cyclase (OSC). The cyclised product is then decorated through oxidation by cytochrome P450s or acetylation by acetyltransferases. The differing combinations of cyclases and tailoring enzymes drives the remarkable diversity of terpene products.

In most studies, the definition of genomic neighbourhoods is arbitrary, i.e. set to a reasonable, but static value (e.g. selecting five genes either side of the gene of interest). This is analogous to antiSMASH (Blin et al. 2019) prediction where BGC boundaries are set a fixed distance up and downstream of the core genes. We decided to take advantage of `cblaster`'s ability to rapidly recompute outputs to define genomic neighbourhoods in a more robust fashion.

To identify triterpine BGCs, we took the sequences of proteins involved in thalianol biosynthesis in *Arabadopsis thaliana* and searched them against the NCBI database. As described in Field et al. (2011), this included a thalianol synthase (At5g48010), two cytochrome P450s (At5g48000 and At5g47990) and an acyltransferase (At5g47980). We then used the `gne` module to iteratively predict BGCs based on increasing intergenic distance values. From this data, we can readily identify suitable values for the intergenic distance thresholds. Intergenic distances below 50 kb are unsuitable, as they are within the 'liminal' region of the curve. Notably, the default

threshold of 20 kb lies well within this region, indicating that many BGCs would be missed by `cblaster` if this parameter were not changed. Values above 475 kb leads to a massive spike in the mean cluster size while cluster count remains stable, indicating that homologues are being hit from more distant regions of the genome. Between these two values, there are a number of stable values suitable for analysis, from a conservative threshold of 50 kb to a more liberal threshold of 120 kb (Figure 4b). As such, `gne` allows the user to rapidly assess suitable intergenic distance values that avoid liminal regions where small variability would have a large effect on the output.

Case Study 3: Identification of biosynthetic gene clusters in fungi

Previously, we reported the *bua* biosynthetic gene cluster (BGC), responsible for the production of the burnettramic acids in *Aspergillus burnettii* (Li et al. 2019). Due to a fragmented genome assembly, initially we could only identify a short fragment of *bua* containing the genes *buaA* and *buaE*, encoding a hybrid polyketide synthase-nonribosomal peptide synthetase (PKS-NRPS) and proline hydroxylase, respectively. As BGCs are often conserved across multiple species, comparative genomics can be used in this scenario to identify the full set of biosynthetic genes (Cacho et al. 2015). Thus, we employed `cblaster` to search for co-located *buaA/E* sequence homologs across publically available genomes in the NCBI database. This resulted in the identification of complete homologous BGCs in other *Aspergillus* spp., including *Aspergillus alliaceus*, which allowed us to map back to other truncated genomic scaffolds in the *A. burnettii genome* and reconstruct the full *bua* BGC (Figure 4c), which we then experimentally characterised.

In another study, we used `cblaster` to identify BGCs encoding the biosynthesis of natural products with 1-benzazepine scaffolds in *Aspergillus* spp. using genes from the nanangelenin biosynthetic pathway (Li et al. 2020). Searching for collacated homologues of the bimodular NRPS, NanA, and the indoleamine-2,3-dioxygenase, NanC, revealed related BGCs in six *Aspergillus* species. In addition, the different clusters contained different complements of tailoring enzymes, such as methyltransferases and P450s, indicating that these BGCs represented novel analogues of the nanangelenin family.

More recently, we reported *hkm*, the BGC encoding the biosynthesis of the hancockiamides, a family of phenylpropanoid piperazines isolated from *Aspergillus hancockii* (Li et al. 2021). While the BGC itself was identified through manual BLASTp searches of the *A. hancockii* genome, using `cblaster` enabled us to identify homologous gene clusters in several other fungal species.

Case Study 4: Identifying drimane sesquiterpenoid BGCs through profile searches

We recently investigated a family of drimane sesquiterpenoids, the nanangenines, isolated from *Aspergillus nanangensis* from section *Jani* (Lacey et al. 2019). Previous work by Shinohara et al. (2016) identified a drimane synthase, AstC, involved in the biosynthesis of the astellolides, a family of drimane sesquiterpenoids. AstC is a novel member of the terpene synthase family and shows similary to haloacid dehydrogenase (HAD)-like hydrolases. The presence of acyl side chains in the nanangenines led us to hypothesize that their biosynthesis would require a drimane synthase, similar to AstC, and either a fatty-acid synthase or PKS. In the study, we

Table 2: Feature comparison of software tools for gene cluster homology searches.

Feature	**cblaster**	**MultiGeneBlast**	**clusterTools**
Query input	Proteins, HMM profiles	Proteins, nucleotides	HMM profiles
Searches against remote databases	X		
Genomic neighbourhood estimation	X		
Extraction of hit sequences	X		
Interactive visualisations	X	X	
Allows input of personal sequence data	X	X	X
Allows creation of local search databases	X	X	X
Graphical user interface available	X	X	X

performed BLAST searches to identify AstC homologues in the *A. nanangensis* genome, and manually filtered the results for only those found collocated with a PKS or FAS. This led to the identification of the candidate cluster (GenBank accession MT024570.1) that we presented in the paper, containing a PKS (FE257_006541) and AstC homologue (FE257_006542).

`cblaster` provides a search mode, `hmm`, which allows for searches against local sequence databases using functional domain profiles from the Pfam database (Mistry et al. 2021). Our functional hypothesis for the biosynthesis of the nanangenines presented a perfect use case for this search mode. To demonstrate, we attempted to directly identify the same candidate gene cluster in the genome using using the domain profiles for a beta-ketoacyl synthase domain (Pfam accession PF00109) and HAD-like hydrolase (Pfam accession PF13419). This search resulted in just two cluster hits, including our manually identified candidate cluster.

We then performed comparative genomics analysis to further support the validity of our candidate cluster. Though drimane sesquiterpenoids are produced by several Aspergilli, those with acyl side chains are unique to species in section *Usti*. We used `cblaster` to search for homologues of our candidate cluster, both remotely in the NCBI BLAST databases, as well as locally against section *Usti* species with genomes deposited in other sources. As a result, we found homologues spanning up to six genes in *Aspergillus ustus*, *Aspergillus calidoustus*, *Aspergillus insuetus* and *Aspergillus pseudodeflectus*, all of which are members of section *Usti* as well as known producers of acyl drimane sesquiterpenoids. Further analyses of these clusters using clinker (Gilchrist and Chooi 2021) gave us confidence that our candidate BGC did indeed encode the biosynthesis of the nanangenines.

Though further studies are necessary to fully characterise this biosynthetic pathway, it shows how `cblaster` has quickly become an indispensable part of this research workflow.

New approaches and practical applications

The case studies above demonstrate that `cblaster` is a powerful tool for identifying homologous gene clusters from the NCBI or local databases. It is possible to query known gene clusters to expand existing families, or user-specified combinations of genes that are predicted to function together. Currently, the most widely used tools for conducting gene cluster homology searches, MultiGeneBlast (Medema et al. 2013) and clusterTools (Lorenzo de los Santos and Challis 2019), depend on local sequence databases. `cblaster` allows remote searching of NCBI sequence databases, eliminating this requirement and reducing the computational burden (Table 2).

`cblaster` provides outputs that can be easily adapted for downstream analyses. Delimited formats can be accessed as spreadsheets, making them accessible to all researchers regardless of bioinformatics experience. The binary table can be sorted and filtered post-analysis, which is particularly useful for narrowing large datasets to focus on individual genes or organisms of interest. In this way, data can be easily extracted to generate simple graphs (e.g. generic distribution, ubiquity of different homologues or average cluster size), build distance matrices and dendrograms. Results tables from multiple searches can easily be concatenated to combine results from local and remote searches. `cblaster extract` can be used to quickly obtain specific homologue sequences for downstream phylogenetic and structural analyses.

By using heatmaps as opposed to more complex gene diagrams, information about the diversity and conservation of encoded proteins is accessible at a glance (Figure 3a). Grouping gene clusters by identity highlights blocks of homologous proteins, facilitating the designation of cluster boundaries and identification of core biosynthetic genes and functional sub-clusters. This is a useful proxy for functional and evolutionary relationships, as shown in Case Study 1 above. In this manner, `cblaster` can uncover deep-seated evolutionary relationships that may not be immediately apparent, such as proteins or sub-clusters that have been co-opted between different pathways. Furthermore, `cblaster` can generate comparative gene cluster visualisations using `clinker` (Gilchrist and Chooi 2021). These visualisations show the detected clusters in their genomic context, enabling a fuller view of the overall conservation of cluster architecture across taxa.

`cblaster gne` can estimate the size of genomic neighbourhoods to help evaluate the robustness of cluster prediction (Figure 3b). This enables the user to identify suitable intergenic distance values that limit missing data without inadvertently introducing false positives. In addition, recent work has highlighted the importance of genomic neighbourhoods in the evolution of gene clusters (Liu et al. 2020b). The distribution of genomic neighbourhood sizes, as calculated by `gne`, provides insight into the conservation of gene cluster topology; a broader distribution indicates a more dynamic architecture. This aspect is likely to be of particular interest in the study of plants and other eukaryotes where gene clusters are more variable in gene density, displayed in Case Study 2 above.

In summary, `cblaster` can uncover homologous gene clusters and evolutionary relationships at a pace that far exceeds other available software and has become an indispensable tool in our own research into the biosynthetic gene clusters of fungi and bacteria. It has already aided us in the identification of several metabolic gene clusters from *Aspergillus* species as discussed in Case Studies 3 and 4 above. We foresee `cblaster` being used to uncover novel variants of known pathways, or to prospect for new combinations of biological parts for use in synthetic biology (Wang et al. 2013; Sun et al. 2015; Gilchrist et al. 2018). `cblaster`'s versatility means it can be applied to any project that requires the study of two or more collocated genes and it is accessible to all researchers regardless of their experience with bioinformatics.

Software implementation and availability

`cblaster` is implemented in Python, with visualisations written in JavaScript and HTML. Source code is freely distributed on GitHub (github.com/gamcil/cblaster) under the MIT license. A comprehensive usage guide as well as API documentation is available online (cblaster.readthedocs.io/en/latest).

Acknowledgements

C.L.M.G. is supported by the Australian Government Research Training Program Ph.D. scholarship. Y-H.C is supported by an Australian Research Council Future Fellowship (FT160100233). This work was funded in part by the Cooperative Research Centres Projects scheme (CRCPFIVE000119).

References

Kai Blin, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*, 47(W1):W81–W87, July 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz310.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3 Data-Driven Documents. *IEEE T Vis Comput Gr*, 17(12):2301–2309, December 2011. ISSN 1077-2626. doi: 10.1109/TVCG.2011.185.

Rosie E. Bradshaw, Jason C. Slot, Geromy G. Moore, Pranav Chettri, Pierre J. G. M. de Wit, Kenneth C. Ehrlich, Austen R. D. Ganley, Malin A. Olson, Antonis Rokas, Ignazio Carbone, and Murray P. Cox. Fragmentation of an aflatoxin-like gene cluster in a forest pathogen. *New Phytologist*, 198(2):525–535, 2013. ISSN 1469-8137. doi: 10.1111/nph.12161.

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, 12(1):59–60, jan 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3176.

Ralph A. Cacho, Yi Tang, and Yit-Heng Heng Chooi. Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. *Front Microbiol*, 5(JAN):1–16, jan 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00774.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10:1–9, 2009. ISSN 14712105. doi: 10.1186/1471-2105-10-421.

Marc G. Chevrette, Karina Gutiérrez-García, Nelly Selem-Mojica, César Aguilar-Martínez, Alan Yañez-Olvera, Hilda E. Ramos-Aboites, Paul A. Hoskisson, and Francisco Barona-Gómez. Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat Prod Rep*, 37(4):566–599, April 2020. ISSN 1460-4752. doi: 10.1039/C9NP00048H.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp163.

Ben Field, Anna Sophie Fiston-Lavier, Ariane Kemen, Katrin Geisler, Hadi Quesneville, and Anne E. Osbourn. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *P Natl Acad Sci USA*, 108 (38):16116–16121, oct 2011. ISSN 00278424. doi: 10.1073/pnas.1109273108.

Fatima Foflonker and Crysten E Blaby-Haas. Colocality to Cofunctionality: Eukaryotic Gene Neighborhoods as a Resource for Function Discovery. *Molecular Biology and Evolution*, 38(2):650–662, February 2021. ISSN 0737-4038. doi: 10.1093/molbev/msaa221.

Danica Galonić Fujimori, Siniša Hrvatin, Christopher S. Neumann, Matthias Strieker, Mohamed A. Marahiel, and Christopher T. Walsh. Cloning and characterization of the biosynthetic gene cluster for kutznerides. *P Natl Acad Sci USA*, 104(42):16498–16503, oct 2007. ISSN 00278424. doi: 10.1073/pnas.0708242104.

Cameron L M Gilchrist and Yit-Heng Chooi. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, (btab007), January 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab007.

Cameron L. M. Gilchrist, Hang Li, and Yit-Heng Chooi. Panning for gold in mould: Can we increase the odds for fungal genome mining? *Org Biomol Chem*, 16(10):1620–1626, 2018. ISSN 1477-0520, 1477-0539. doi: 10.1039/C7OB03127K.

Wun Rong Hong, Mei Ge, Zhi Hong Zeng, Li Zhu, Min Yu Luo, Lei Shao, and Dai Jie Chen. Molecular cloning and sequence analysis of the sisomicin biosynthetic gene cluster from Micromonospora inyoensis. *Biotechnol Lett*, 31(3):449–455, 2009. ISSN 15736776. doi: 10.1007/s10529-008-9887-y.

Chuan Huang, Fanglu Huang, Eileen Moison, Junhong Guo, Xinyun Jian, Xiaobo Duan, Zixin Deng, Peter F. Leadlay, and Yuhui Sun. Delineating the biosynthesis of gentamicin X2, the common precursor of the gentamicin C antibiotic complex. *Chem Biol*, 22(2):251–261, feb 2015. ISSN 10745521. doi: 10.1016/j.chembiol.2014.12.012.

Satria A Kautsar, Kai Blin, Simon Shaw, Jorge C Navarro-Muñoz, Barbara R Terlouw, Justin J J van der Hooft, Jeffrey A van Santen, Vittorio Tracanna, Hernando G Suarez Duran, Victòria Pascal Andreu, Nelly Selem-Mojica, Mohammad Alanjary, Serina L Robinson, George Lund, Samuel C Epstein, Ashley C Sisto, Louise K Charkoudian, Jérôme Collemare, Roger G Linington, Tilmann Weber, and Marnix H Medema. MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res*, page gkz882, October 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkz882.

Simon C. Kessler, Xianghui Zhang, Megan C. McDonald, Cameron L. M. Gilchrist, Zeran Lin, Adriana Rightmyer, Peter S. Solomon, B. Gillian Turgeon, and Yit-Heng Chooi. Victorin, the host-selective cyclic

peptide toxin from the oat pathogen *Cochliobolus victoriae* , is ribosomally encoded. *Proceedings of the National Academy of Sciences*, 117(39):24243–24250, September 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2010573117.

Mamoru Komatsu, Muneya Tsuda, Satoshi Ōmura, Hideaki Oikawa, and Haruo Ikeda. Identification and functional analysis of genes controlling biosynthesis of 2-methylisoborneol. *Proceedings of the National Academy of Sciences*, 105(21):7422–7427, May 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0802312105.

Heather J. Lacey, Cameron L. M. Gilchrist, Andrew Crombie, John A. Kalaitzis, Daniel Vuong, Peter J. Rutledge, Peter Turner, John I. Pitt, Ernest Lacey, Yit-Heng Chooi, and Andrew M. Piggott. Nanangenines: Drimane sesquiterpenoids as the dominant metabolite cohort of a novel Australian fungus, Aspergillus nanangensis. *Beilstein J Org Chem*, 15(1):2631–2643, November 2019. ISSN 1860-5397. doi: 10.3762/bjoc.15.256.

J. M. Lee. Genomic Gene Clustering Analysis of Pathways in Eukaryotes. *Genome Research*, 13(5):875–882, May 2003. ISSN 10889051. doi: 10.1101/gr.737703.

Hang Li, Cameron L.M. Gilchrist, Heather J. Lacey, Andrew Crombie, Daniel Vuong, John I. Pitt, Ernest Lacey, Yit Heng Chooi, and Andrew M. Piggott. Discovery and Heterologous Biosynthesis of the Burnettramic Acids: Rare PKS-NRPS-Derived Bolaamphiphilic Pyrrolizidinediones from an Australian Fungus, Aspergillus burnettii. *Org Lett*, 21(5):1287–1291, 2019. ISSN 15237052. doi: 10.1021/acs.orglett.8b04042.

Hang Li, Cameron L.M. Gilchrist, Chin Soon Phan, Heather J. Lacey, Daniel Vuong, Stephen A. Moggach, Ernest Lacey, Andrew M. Piggott, and Yit Heng Chooi. Biosynthesis of a New Benzazepine Alkaloid Nanangelenin A from Aspergillus nanangensis Involves an Unusual l-Kynurenine-Incorporating NRPS Catalyzing Regioselective Lactamization. *J Am Chem Soc*, 142(15):7145–7152, apr 2020. ISSN 15205126. doi: 10.1021/jacs.0c01605.

Hang Li, Alastair E. Lacey, Si Shu, John A. Kalaitzis, Daniel Vuong, Andrew Crombie, Jinyu Hu, Cameron L. M. Gilchrist, Ernest Lacey, Andrew M. Piggott, and Yit-Heng Chooi. Hancockiamides: Phenylpropanoid piperazines from Aspergillus hancockii are biosynthesised by a versatile dual single-module NRPS pathway. *Organic & Biomolecular Chemistry*, 19(3):587–595, 2021. doi: 10.1039/D0OB02243H.

Zhenhua Liu, Jitender Cheema, Marielle Vigouroux, Lionel Hill, James Reed, Pirita Paajanen, Levi Yant, and Anne Osbourn. Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nature Communications*, 11(1):5354, October 2020a. ISSN 2041-1723. doi: 10.1038/s41467-020-19153-6.

Zhenhua Liu, Hernando G. Suarez Duran, Yosapol Harnvanichvech, Michael J. Stephenson, M. Eric Schranz, David Nelson, Marnix H. Medema, and Anne Osbourn. Drivers of metabolic diversification: How dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. *New Phytol*, 227(4):1109–1123, 2020b. ISSN 1469-8137. doi: 10.1111/nph.16338.

Zhenhua Liu, Hernando G. Suarez Duran, Yosapol Harnvanichvech, Michael J. Stephenson, M. Eric Schranz, David Nelson, Marnix H. Medema, and Anne Osbourn. Drivers of metabolic diversification: how dynamic

genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. *New Phytol*, 227(4):1109–1123, aug 2020c. ISSN 0028-646X. doi: 10.1111/nph.16338.

Emmanuel Lorenzo de los Santos and Gregory Challis. clusterTools: Functional element identification for the in silico prioritization of biosynthetic gene clusters. *Access Microbiol*, 1(1A), 2019. doi: 10.1099/acmi.ac2019.po0154.

Marnix H. Medema, Eriko Takano, and Rainer Breitling. Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. *Mol Biol Evol*, 30(5):1218–1223, May 2013. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/mst025.

Pawel Michalak. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3):243–248, March 2008. ISSN 0888-7543. doi: 10.1016/j.ygeno.2007.11.002.

Matej Mihelčić, Tomislav Šmuc, and Fran Supek. Patterns of diverse gene functions in genomic neighborhoods predict gene function and phenotype. *bioRxiv*, page 582577, mar 2019. doi: 10.1101/582577.

Jaina Mistry, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12):e121–e121, July 2013. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkt263.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa913.

Jorge C. Navarro-Muñoz, Nelly Selem-Mojica, Michael W. Mullowney, Satria A. Kautsar, James H. Tryon, Elizabeth I. Parkinson, Emmanuel L. C. De Los Santos, Marley Yeong, Pablo Cruz-Morales, Sahar Abubucker, Arne Roeters, Wouter Lokhorst, Antonio Fernandez-Guerra, Luciana Teresa Dias Cappelini, Anthony W. Goering, Regan J. Thomson, William W. Metcalf, Neil L. Kelleher, Francisco Barona-Gomez, and Marnix H. Medema. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology*, 16(1):60–68, January 2020. ISSN 1552-4469. doi: 10.1038/s41589-019-0400-9.

NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 45(D1):D12–D17, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1071.

Hans Wilhelm Nützmann, Ancheng Huang, and Anne Osbourn. Plant metabolic clusters – from genetics to genomics. *New Phytol*, 211(3):771–789, aug 2016. ISSN 14698137. doi: 10.1111/nph.13981.

Hans-Wilhelm Nützmann, Claudio Scazzocchio, and Anne Osbourn. Metabolic Gene Clusters in Eukaryotes. *Annu Rev of Genet*, 52(1):159–183, November 2018. ISSN 0066-4197, 1545-2948. doi: 10.1146/annurev-genet-120417-031237.

César Sánchez, Igor A. Butovich, Alfredo F. Braña, Jürgen Rohr, Carmen Méndez, and José A. Salas. The Biosynthetic gene cluster for the antitumor rebeccamycin: Characterization and generation of indolocarbazole derivatives. *Chem Biol*, 9(4):519–531, 2002. ISSN 10745521. doi: 10.1016/S1074-5521(02)00126-6.

Yasutomo Shinohara, Shunji Takahashi, Hiroyuki Osada, and Yasuji Koyama. Identification of a novel sesquiterpene biosynthetic machinery involved in astellolide biosynthesis. *Scientific Reports*, 6(1):32865, December 2016. ISSN 2045-2322. doi: 10.1038/srep32865.

Sangkeun Son, Young-Soo Hong, Mina Jang, Kyung Taek Heo, Byeongsan Lee, Jun-Pil Jang, Jong-Won Kim, In-Ja Ryoo, Won-Gon Kim, Sung-Kyun Ko, Bo Yeon Kim, Jae-Hyuk Jang, and Jong Seog Ahn. Genomics-Driven Discovery of Chlorinated Cyclic Hexapeptides Ulleungmycins A and B from a Streptomyces Species. *J Nat Prod*, 80(11):3025–3031, nov 2017. ISSN 1520-6025. doi: 10.1021/acs.jnatprod.7b00660.

Huihua Sun, Zihe Liu, Huimin Zhao, and Ee Lui Ang. Recent advances in combinatorial biosynthesis for drug discovery, feb 2015. ISSN 11778881.

Jamie Unwin, Scott Standage, Dylan Alexander, Thomas Hosted, Ann C. Horan, and Elizabeth M.H. Wellington. Gene cluster in Micromonospora echinospora ATCC15835 for the biosynthesis of the gentamicin C complex. *J Antibiot*, 57(7):436–445, 2004. ISSN 00218820. doi: 10.7164/antibiotics.57.436.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, Yoshiki Vázquez-Baeza, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods*, 17(3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2.

Shengqin Wang, Zhihong Zheng, Huixi Zou, Nan Li, and Mingjiang Wu. Characterization of the secondary metabolite biosynthetic gene clusters in archaea. *Comput Biol and Chem*, 78:165–169, February 2019. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2018.11.019.

Yaya Wang, Jiali Wang, Shuqi Yu, Fan Wang, Hongmin Ma, Changwu Yue, Minghao Liu, Zixin Deng, Ying Huang, and Xudong Qu. Identifying the Minimal Enzymes for Unusual Carbon-Sulfur Bond Formation in Thienodolin Biosynthesis. *ChemBioChem*, 17(9):799–803, may 2016. ISSN 14397633. doi: 10.1002/cbic.201500670.

Yen Hsiang Wang, Kathy Y. Wei, and Christina D. Smolke. Synthetic biology: Advancing the design of diverse genetic systems. *Annu Rev Chem Biomol*, 4:69–102, 2013. ISSN 19475438. doi: 10.1146/annurev-chembioeng-061312-103351.

P. Wiemann, C.-J. Guo, J. M. Palmer, R. Sekonyela, C. C. C. Wang, and N. P. Keller. Prototype of an intertwined secondary-metabolite supercluster. *Proceedings of the National Academy of Sciences*, 110(42):17065–17070, October 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1313258110.