

DNA and Protein Motif Pattern searches in FungiDB

This tutorial uses <https://beta.fungidb.org>

Example 1. Identify GATA motifs in *Aspergillus fumigatus* Af293 using DNA Motif Pattern Search

There are three major classes of eukaryotic zinc finger proteins. For example, Class II represents a group of proteins that includes transcription factors binding to GATA DNA motif. This motif occurs in the regulatory regions of the target genes and can be defined as [(A/T)GATA(A/G)] (A or T, followed by GATA, followed by A or G).

Using the VEuPathDB Regular Expression Tutorial, we can re-write the motif as follows: [AT]GATA[AG], where [] means “match any character contained in the brackets”.

To deploy a search for GATA DNA motif in *Aspergillus fumigatus* Af293, follow these steps:

- Navigate to the **Search for...** and filter for **motif**,
- Select **DNA Motif Pattern**,
- Find *A. fumigatus* Af293 genome by filtering genomes for **fumig**,
- Enter [AT]GATA[AG] in the **Pattern** parameter and click **Get Answer**

The screenshot shows the FungiDB search interface. On the left, the 'Search for...' panel has a search bar with 'motif' entered. Below the search bar, there are buttons for 'Genes', 'Sequence analysis', 'Protein Motif Pattern', 'Genomic Segments', and 'DNA Motif Pattern'. An arrow points from the 'DNA Motif Pattern' button to the right panel. The right panel, titled 'Identify Genomic Segments based on DNA Motif Pattern', has two sections: 'Organism' and 'Pattern'. In the 'Organism' section, 'fumig' is selected, and a list of organisms is shown, including 'Aspergillus fumigatus Af293'. In the 'Pattern' section, '[AT]GATA[AG]' is entered.

59465 Genomic Segments from Step 1 [Revise](#)

Strategy: DNA Motif(9)

Genomic Segment Results [Genomic Locations](#)

Rows per page: 1000

[Download](#) [Add to Basket](#) [Add Columns](#)

Segment ID	Organism	Genomic Location	Motif	Length
Chr1_A_fumigatus_Af293:1000359-1000365:r	Aspergillus fumigatus Af293	Chr1_A_fumigatus_Af293:1,000,359..1,000,365 (-)	...CGCAGGGATAAACTACAGGGAGATAGTGCTACTTTCATGTCAGAGT...	7
Chr1_A_fumigatus_Af293:1000521-1000527:f	Aspergillus fumigatus Af293	Chr1_A_fumigatus_Af293:1,000,521..1,000,527 (+)	...TCAAGGTCACATTACGTTGTTGATAGAAATAACCCCAATTTTCTGT...	7
Chr1_A_fumigatus_Af293:1001672-1001678:r	Aspergillus fumigatus Af293	Chr1_A_fumigatus_Af293:1,001,672..1,001,678 (-)	...ACGATAGTCGCGCTATCAGTGATAGTGTCGCGGTTGCGAACTCTA...	7
Chr1_A_fumigatus_Af293:1001679-1001685:f	Aspergillus fumigatus Af293	Chr1_A_fumigatus_Af293:1,001,679..1,001,685 (+)	...CGCAACCGCGACATTATCACTGATAGCGGCGACTATCGTCGTGGAG...	7
Chr1_A_fumigatus_Af293:1001866-1001872:r	Aspergillus fumigatus Af293	Chr1_A_fumigatus_Af293:1,001,866..1,001,872 (-)	...GGTGGCAAGACACCAATATCTGATAGCACTGCTCGCGCTCCCGCTG...	7

Further reading: <https://mmb.r.asm.org/content/70/3/583>

Example 2. Find fungal genes downstream of a regulatory DNA motif.

DNA motifs are important for the regulation of processes linked to host cell invasion, production of secondary metabolites, etc. The basic-helix-loop-helix (b-HLH) motif is involved in transcriptional regulation and cell-type determination. *S. cerevisiae* transcription factor PHO4 is one of the b-HLH motif proteins. It binds to the CACGTG DNA motif and positively regulates the acid phosphatase *PHO5*.

Identify all genes with upstream (600bp) CACGTG DNA motif in *S. cerevisiae* S288c and determine if PHO5 is one of those genes.

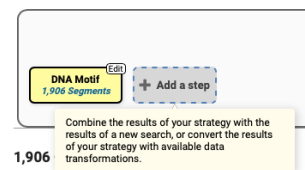
- Navigate to the **Search for** and filter for **motif**,
- Select **DNA Motif Pattern**, which is located in the **Genomic Segments** menu;
- Find *S. cerevisiae* genome by filtering genomes for **cerev**
- Enter CACGTG in the Pattern parameter and click **Get Answer**

The image shows two parts of a web interface. On the left, a 'Search for...' sidebar with a search bar containing 'motif' and a dropdown menu with 'Genes', 'Sequence analysis' (with a sub-item 'Protein Motif Pattern'), and 'Genomic Segments' (with a sub-item 'DNA Motif Pattern'). A pink arrow points from the 'DNA Motif Pattern' option to a larger form on the right. The form is titled 'Identify Genomic Segments based on DNA Motif Pattern'. It includes a note: 'Note: You may select up to 1 values for this parameter. 1 selected, out of 164'. Below this is a search bar with 'cerev' and a dropdown menu with 'Fungi', 'Ascomycota', 'Saccharomycetes', 'Saccharomycetales', 'Saccharomycetaceae', 'Saccharomyces', 'Saccharomyces cerevisiae', and 'Saccharomyces cerevisiae S288c'. Below the dropdown is a 'Pattern' section with a text input field containing 'CACGTG' and a 'Get Answer' button.

The image shows a 'My Search Strategies' page. At the top, it says 'Opened (1) All (963) Public (37) Help'. Below this is a section for 'Unnamed Search Strategy' with a 'DNA Motif' button and an 'Add a step' button. Below this is a section for '1,906 Genomic Segments' with a 'Revise this search' button. Below this is a table with the following columns: Segment ID, Organism, Genomic Location, Motif, and Length. The table contains 4 rows of data.

Segment ID	Organism	Genomic Location	Motif	Length
BK006934:1958-1964.f	Saccharomyces cerevisiae S288c	BK006934:1,958..1,964 (+)	...CATTAGCACTACCATGAATGCACGTGTCGCTGCTCATCACTGCT...	7
BK006934:1958-1964.r	Saccharomyces cerevisiae S288c	BK006934:1,958..1,964 (-)	...AGCAGTGATGAGGACAGCAGCACGTGCATTTCATGCTAGTCTAATG...	7
BK006934:107613-107619.f	Saccharomyces cerevisiae S288c	BK006934:107,613..107,619 (+)	...ATTGCAAGTAGTATTTTGTACGTGATTTTGATCCAATTAACACT...	7
BK006934:107613-107619.r	Saccharomyces cerevisiae S288c	BK006934:107,613..107,619 (-)	...AGTTTAAATTGGATCAAAATCACGTGACAAAATACTACTTTGCAAT...	7

- Next, click on the **+ Add a step** button



- Select **Use Genomic Colocation... > A new Search > Genes**

Add a step to your search strategy

Use the relative position of features on the genome between your existing step and the new step to identify features to keep in the final result.

Choose *which* features to collocate. From...

☒ A new search ☐ An existing strategy ☐ My basket

expand all | collapse all

Filter the searches below...

- Genes
- Genomic Segments
- SNPs

Combine with other Genomic Segments

DNA Motif
1,906 Segments
Step 1

Step 2

Use Genomic Colocation to combine with other features

DNA Motif
1,906 Segments
Step 1

Step 2

- Expand **Genes** options by clicking and select **Taxonomy**, and then **Organism**.

Choose which features to collocate. From...

Use the relative position of features on the genome between your existing step and the new step to identify features to keep in the final result.

Choose which features to collocate. From...

☒ A new search ☐ An existing strategy ☐ My basket

expand all | collapse all

Filter the searches below...

- Genes
 - Annotation, curation and identifiers
 - Function prediction
 - Gene models
 - Genetic variation
 - Genomic Location
 - Immunology
 - Orthology and synteny
 - Pathways and interactions
 - Phenotype
 - Protein features and properties
 - Protein targeting and localization
 - Proteomics
 - Sequence analysis
 - Structure analysis
 - Taxonomy
 - Text
 - Transcriptomics
- Genomic Segments
- SNPs

Organism

- Filter on **cerev** to select **S. cerevisiae** genome and click on the **Continue...** button

← Add a step to your search strategy ⓘ

🔍 Organism

1 selected, out of 148

[add these](#) | [clear these](#) | [select only these](#)
[select all](#) | [clear all](#)

cerev x ⓘ

- 📁 Fungi
 - 📁 Ascomycota
 - 📁 Saccharomycetes
 - 📁 Saccharomycetaceae
 - 📁 Saccharomyces
 - 📁 Saccharomyces cerevisiae S288c

[add these](#) | [clear these](#) | [select only these](#)
[select all](#) | [clear all](#)

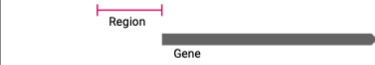
Continue...

🔍 Give this search a name (optional)
 🔍 Give this search a weight (optional)

- Set up colocation tool to ***Return each Gene from the new step*** whose ***upstream 600 bp region overlaps*** the ***exact*** region of a Genomic Segment (which is the CACGTG motif) from the current step and is on either strand.

← Add a step to your search strategy ⓘ

"Return each whose **upstream region** overlaps of a Genomic Segment from the current step and is on



☐ Exact
☒ Upstream: bp
☐ Downstream: bp
☐ Custom:
 begin at: start bp
 end at: start bp



☒ Exact
☐ Upstream: bp
☐ Downstream: bp
☐ Custom:
 begin at: start bp
 end at: stop bp

Run Step

- Click ***Run Step***

449 Genes (409 ortholog groups)

Organism Filter
select all | clear all | expand all | collapse all
Hide zero counts
Search organisms...
Fungi 449
Oomycetes 0
select all | clear all | expand all | collapse all
Hide zero counts

Gene Results | Genome View | **Analyze Results**

Rows per page: 1000

Download | Add to Basket | Add Columns

	Gene ID	Transcript ID	Organism	Genomic Location (Transcript)	Product Description	Match Count	Region
	YHL038C	YHL038C-126_1	<i>Saccharomyces cerevisiae S288c</i>	BK006934:23617..25509(-)	Cbp2p	2	25510 - 26109 (-)
	YHL036W	YHL036W-126_1	<i>Saccharomyces cerevisiae S288c</i>	BK006934:26241..27881(+)	Mup3p	2	25641 - 26240 (+)

- To check if *PHO5* is one of the 449 genes returned, click on the **+ Add a step** button, choose to **Combine with other Genes**, and **intersect** with **Text** search:

Add a step to your search strategy

Combine with other Genes

1 Choose how to combine with other Genes
☒ 2 INTERSECT 3
☐ 2 UNION 3
☐ 2 MINUS 3
☐ 3 MINUS 2

2 Choose which Genes to combine. From...
☒ A new search
☐ An existing strategy
☐ My basket

Text
 Text (product name, notes, etc.)
 Find genes with a text search against their product name, notes, GO, EC, Domains, NRDB, or metabolic pathways.

- Search for a text term *PHO5* in **Gene name or symbol**:

Add a step to your search strategy

cer

Fungi
 Ascomycota
 Saccharomycetes
 Saccharomycetales
 Saccharomycetaceae
 Saccharomyces
 Saccharomyces cerevisiae
 Saccharomyces cerevisiae S288c

add these | clear these | select only these
 select all | clear all

Text term (use * as wildcard)

PHO5

Fields

☐ Alternate product descriptions
☐ EC descriptions and numbers
☐ Epitopes from IEDB
☐ External links
☐ Gene ID
☒ Gene name or symbol
☐ Gene type

My Search Strategies

Opened (1) All (963) Public (37) Help

Unnamed Search Strategy *

1 Gene (1 ortholog groups)

Gene Results | Genome View | **Analyze Results**

Rows per page: 1000

Download Add to Basket Add Columns

Gene ID	Transcript ID	Product Description	Genomic Location (Gene)	Gene Name or Symbol
YBR093C	YBR093C-126_1	acid phosphatase PHO5	BK006936:429,548..430,951(-)	PHO5

Rows per page: 1000

Organism Filter
select all | clear all | expand all | collapse all
☐ Hide zero counts
Search organisms...
Fungi 1
Oomycetes 0
select all | clear all | expand all | collapse all
☐ Hide zero counts

Further reading: <https://www.pnas.org/content/99/26/16893>

Example 3. Identify CaaX box proteins using Protein Motif Search

CaaX box proteins undergo several modifications essential to their targeting. These eukaryotic proteins contain a motif where cysteine is followed by the two aliphatic amino acids, followed by an amino acid that confers substrate specificity (e.g. proteins with C-terminal cysteine-aliphatic-aliphatic-M/S/Q/A/C/L/E amino acids, where M/S/Q/A/C/L/E confer substrate specificity).

Using the information below and the VEuPathDB Regular Expression help, define CaaX motif and identify corresponding gene in *S. cerevisiae*.

- Aliphatic amino acids** are non-polar and hydrophobic that are mostly found within protein molecules (e.g. isoleucine, leucine, valine), while alanine and glycine may be found either inside or outside a protein molecule.
- Amino acids codes:**

Cysteine – C	Isoleucine – I	Leucine – L	Valine – V
Alanine – A	Glycine – G	Serine – S	Methionine – M
Glutamine – Q	Glutamic acid - E		
- Use **\$** to match only at the end of the string or C-terminus

The C-terminal CaaX (cysteine-aliphatic-aliphatic-M/S/Q/A/C/L/E) motif can be re-written as:

C[ILVAG][ILVAG][MSQACLE]\$

C - Cysteine

[ILVAG] - matches any aliphatic amino acid (isoleucine, leucine, valine, alanine, glycine)

[MSQACLE] - matches any amino acid that confers substrate specificity

\$ - match only at the end of the string or C-terminus

Deploy a search for CaaX protein motif in *S. cerevisiae* S288c:

- Navigate to the **Search for...** and filter for **motif**;
- Select **Protein Motif Pattern**;
- Find *S. cerevisiae* genome by filtering genomes for **cerev**;
- Enter C[ILVAG][ILVAG][MSQACLE]\$ in the **Pattern** parameter and click **Get Answer**

Search for...

motif

Genes

Sequence analysis

Protein Motif Pattern

Genomic Segments

DNA Motif Pattern

Identify Genes based on Protein Motif Pattern

Pattern

C[AGILV][AGILV][MSQACLE]\$

Organism

1 selected, out of 148

cere

Fungi

Saccharomycetes

Saccharomycetes

Saccharomycetes cerevisiae

Saccharomycetes cerevisiae S288c

add these | clear these | select only these

select all | clear all

My Strategies:

New

Opened (5)

All (1079)

Basket

Public Strategies (66)

Help

Hide search strategy panel

(Genes)

Prot Motif

17 Genes

Step 1

Add Step

Strategy: Prot Motif(10)

Rename

Duplicate

Save As

Share

Delete

17 Genes from Step 1

Revise

Strategy: Prot Motif(10)

Click on a number in this table to limit/filter your results

Gene Results

Genome View

Analyze Results

Rows per page: 1000

Download

Add to Basket

Add Columns

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Match Locations	Match Count	Motif
YBL061C	YBL061C-126_1	S. cerevisiae S288c	BK006936:105,316..107,406(-)	Skt5p	(693-696)	1	... SKKPTSLKNNKDKQGKKKKDCVIM
YDL009C	YDL009C-126_1	S. cerevisiae S288c	BK006938:432,925..433,248(-)	hypothetical protein	(104-107)	1	... AQPRIYHHSRLVMILKVSLECAVS
YDR461W	YDR461W-126_1	S. cerevisiae S288c	BK006938:1,385,176..1,385,286(+)	mating pheromone a	(33-36)	1	MQPSTATAAPKEKTSSEKDKNYIIGVFWDPACV
YGL082W	YGL082W-126_1	S. cerevisiae S288c	BK006941:355,827..356,972(+)	hypothetical protein	(378-381)	1	... KFLPFNGSNKEKKRDKLKKNCVIM

Further reading: <https://ec.asm.org/content/5/9/1560>