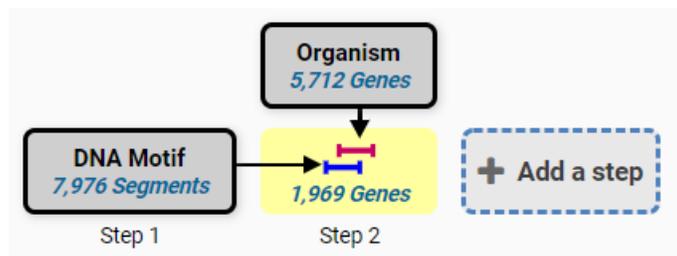


VEuPathDB Webinar 4 June 2020: Motifs and Regular expressions

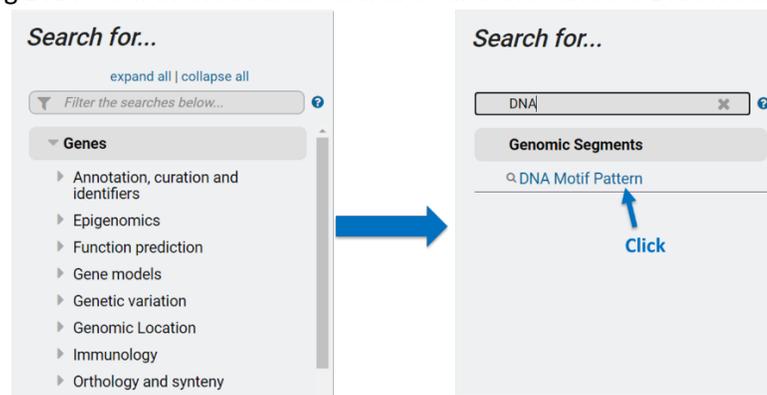
Slide deck - https://docs.google.com/presentation/d/1vvLkSxix-9S5718WB0LLe4wWZ1QD7FtsbEtkKpf5dLU/edit#slide=id.g862fa1b519_3_168

Example: Find EcoR1 restriction enzyme sites within *Plasmodium falciparum* 3D7 genes.

EcoR1 is a restriction enzyme that will cleave DNA whenever it finds the sequence **GAATTC**. EcoR1 can tolerate no ambiguity in the DNA sequence; it always needs GAATTC. VEuPathDB sites have a search called 'Identify Genomic Segments based on DNA Motif Pattern' that will locate and return DNA motifs as genomic segments (a VEuPathDB record type). We will use the DNA Motif Pattern search to find areas of the *Plasmodium falciparum* 3D7 genome that contain the GAATTC motif. The search returns a list of genomic segments. To correlate these genomic segments with genes, we will run a second search for all genes in *Plasmodium falciparum* 3D7 and use the colocation tool to reduce the list to only those genes that contain the genomic segments we found with the DNA Motif Pattern search. The final strategy will look like this. although future releases of PlasmoDB may return different numbers, since the genome and annotation for *Plasmodium falciparum* 3D7 may change.



1. Navigate to the DNA motif search.
 - a. Go to the home page beta.plasmodb.org <https://beta.plasmodb.org/plasmo.beta/app/>
 - b. Begin typing DNA motif in the Search For... filter and then choose DNA Motif Pattern



2. Initiate the DNA Motif Pattern Search
 - a. Choose to search *Plasmodium falciparum* 3D7.
 - b. Type the EcoR1 recognition site, GAATTC, into the **Pattern** field.
 - c. Click **Get Answer**

Identify Genomic Segments based on DNA Motif Pattern

Organism

*Note: You may select up to 1 values for this parameter.
1 selected, out of 45*

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

3d7

- Aconoidasida
 - Haemosporida
 - Plasmodiidae
 - Plasmodium
 - Plasmodium falciparum
 - Plasmodium falciparum 3D7

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

Pattern

GAATTQ

- d. Your results are 'segments' of DNA that contain the motif. Each segment has a known location within the *P. falciparum* 3D7 genome.

My Search Strategies

Opened (1) All (304) Public (44) Help

Unnamed Search Strategy*

DNA Motif
7,976 Segments

Step 1

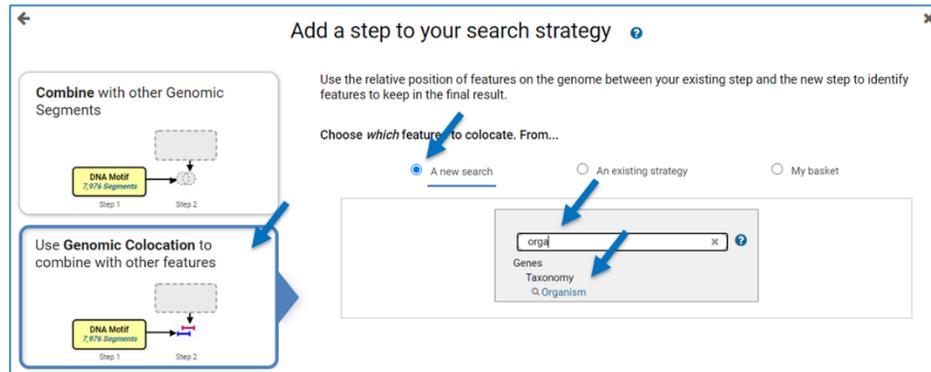
7,976 Genomic Segments

Genomic Segment Results | Genomic Locations

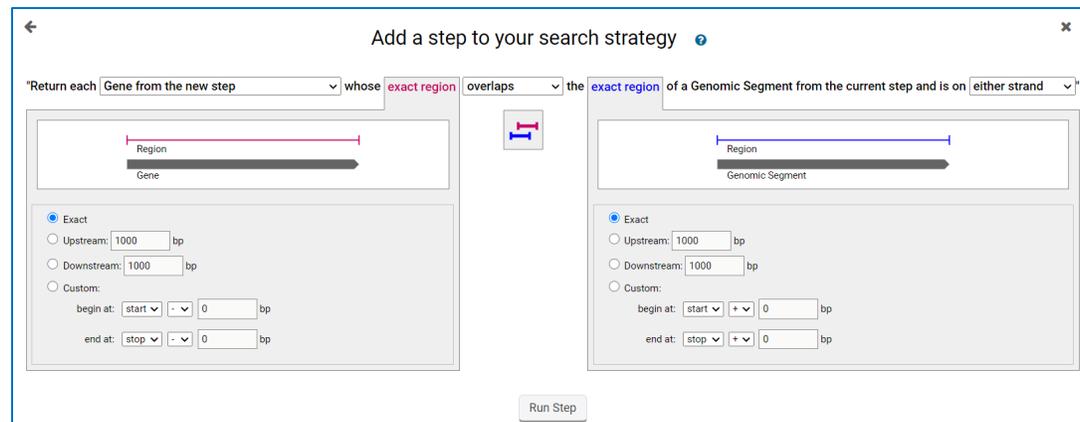
Rows per page: 20

Segment ID	Organism	Genomic Location	Motif
PF3D7_01_v3:101289-101295f	Plasmodium falciparum 3D7	PF3D7_01_v3:101,289..101,295 (+)	...GTAGGAAGTAGTGAACAATGAATCAATAACATACAAATTCGAAA...
PF3D7_01_v3:101289-101295r	Plasmodium falciparum 3D7	PF3D7_01_v3:101,289..101,295 (-)	...TTTCGAAATTGTATGTTATTGAATTCATTTGTTCACTACTCTCTAC...
PF3D7_01_v3:112882-112888f	Plasmodium falciparum 3D7	PF3D7_01_v3:112,882..112,888 (+)	...CTGTTGCTCTATCTACAACAGAATTCATTTGTTTCAGTACTGA...

3. Colocate these EcoR1 motifs with *Plasmodium falciparum* 3D7 genes.
 - a. Click **Add a step** and choose **Use Genomic Colocation** to combine with other features.
 - b. Choose to run **A new search** for **Genes** based on **Taxonomy, Organism**



- c. From the search form choose *P. falciparum* 3D7 and click Get Answer
- d. Arrange the colocation tool to read "Return each gene from the new step whose exact region overlaps the exact region of the genomic segment from the current step and is on either strand".



- e. The result is a list of genes that contain the EcoR1 restriction site (the motif we searched for) and the Region column shows the location of the motif.

Unnamed Search Strategy *

1,969 Genes (1,791 ortholog groups)

Gene Results Genome View Analyze Results

Genes: 1,969 Transcripts: 1,999 Show Only One Transcript Per Gene

Rows per page: 20

Gene ID	Transcript ID	Organism	Genomic Location (Transcript)	Product Description	Match Count	Region	Matched Regions
PF3D7_0100300	PF3D7_0100300.1	<i>Plasmodium falciparum</i> 3D7	PI3D7_01_v3:42367..46507(-)	erythrocyte membrane protein 1, PfEMP1	2	42367 - 46507 (-)	PI3D7_01_v3:44532-44538:r:44,532 - 44,538 (-); PI3D7_01_v3:44532-44538:f:44,532 - 44,538 (+)
PF3D7_0102200	PF3D7_0102200.1	<i>Plasmodium falciparum</i> 3D7	PI3D7_01_v3:98819..102282(+)	ring-infected erythrocyte surface antigen	2	98819 - 102282 (+)	PI3D7_01_v3:101289-101295:r:101,289 - 101,295 (-); PI3D7_01_v3:101289-101295:f:101,289 - 101,295 (+)

Example: Find Avall restriction enzyme sites within *Plasmodium falciparum* 3D7 genes.

We will use the workflow shown above to find Avall sites within *Plasmodium falciparum* 3D7 genes. Avall is a restriction enzyme that will cleave DNA whenever it finds the sequence **GGACC** or **GGTCC**. So Avall can tolerate some ambiguity in the DNA binding sequence; the middle base can either be an A or a T. To inform the 'Identify Genomic Segments based on DNA Motif Pattern' search about the ambiguous motif that we want to find, we must use a regular expression to communicate the motif sequence.

GGACC or GGTCC = GG[AT]CC (regular expression)

The workflow above will find genes that contain the DNA motif in Step 1. The only change needed in the workflow is to enter the Avall regular expression for the Pattern in Step 1.

Revise your step

Identify Genomic Segments based on DNA Motif Pattern

Organism

Note: You may select up to 1 values for this parameter.
1 selected, out of 45

select all | clear all | expand all | collapse all

Filter list below...

- Aconoidasida
- Haemosporida

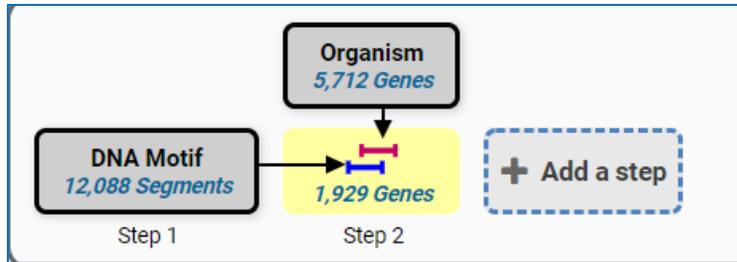
select all | clear all | expand all | collapse all

Pattern **GG[AT]CC**

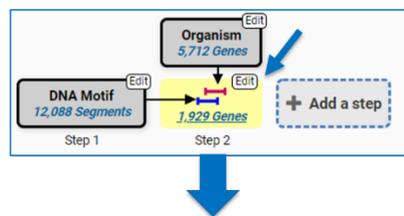
GG[AT]CC

Revise

The final strategy will look like this:



Optional: Find Avall sites upstream of Pf3D7 genes. The above workflow finds genes that contain restriction sites. To find genes with upstream restriction sites, change the colocation tool to read. “Return each Gene from Step 2 whose upstream region (Upstream 1000bp) overlap with the genomic segments in Step 1 and are on either strand.”



View | Analyze | Insert step before | Orthologs | Delete

Details for step *Genes by Rel Loc*

1969 Genes

Return each *Gene from Step 2* whose *upstream region* overlaps the *exact region* of a Genomic Segment from Step 1 and is on *either strand*

Region

Gene

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: start - 1000 bp

end at: start - 1 bp

Region

Genomic Segment

Exact

Upstream: 1000 bp

Downstream: 1000 bp

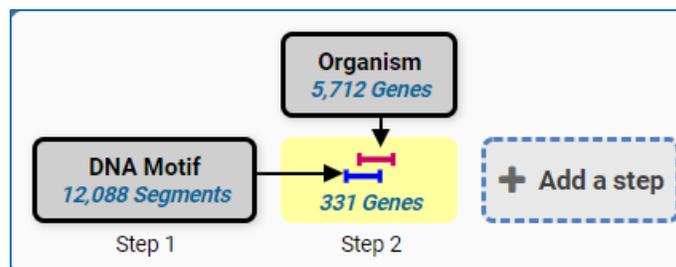
Custom:

begin at: start + 0 bp

end at: stop + 0 bp

Revise

Your result will change to:



Example: Explore *Anopheles gambiae* str. PEST genes that have zinc finger domains and the genes that they may regulate.

A zinc finger is a small protein structural motif that is characterized by the coordination of one or more zinc ions (Zn²⁺) to stabilize the fold. Proteins that contain zinc fingers (zinc finger proteins) are classified into several different structural families. In spite of the large variety of these proteins, the vast majority of zinc finger proteins function as interaction modules that bind DNA, RNA, proteins. Variations in structure serve primarily to alter the binding specificity. (https://en.wikipedia.org/wiki/Zinc_finger)

A zinc finger protein will carry some version of a zinc finger structural motif. Once such motif is described by the PROSITE entry PS00344 GATA-ZN_FINGER_1 <https://prosite.expasy.org/PS00344>. The pattern described for this zinc finger is:

C-x-[DNEHQSTI]-C-x(4,6)-[ST]-x(2)-[WM]-[HR]-[RKENAMSLPGQT]-x(3,4)-[GNEP]-x(3,6)-C-[NES]-[ASNR]-C

Written as a regular expression, the motif is:

C.[DNEHQSTI]C.{4,6}[ST].{2}[WM][HR][RKENAMSLPGQT].{3,4}[GNEP].{3,6}C[NES][ASNR]C

The genes that zinc finger proteins regulate will contain a 'GATA box' DNA binding motif in their regulatory regions. The motif is short and relatively general:

(A/T)GATA(A/G)

and can be written as a regular expression:

[AT]GATA[AG].

In this exercise we will use beta.vectorbase.org to find *Anopheles gambiae* str. PEST genes whose proteins contain our zinc finger domain and we will also search for possible binding partners by searching for genes that contain the GATA zinc finger domain upstream of their start sites.

1. Navigate to the Protein Motif Pattern search.
 - a. Go to the home page beta.vectorbase.org
<https://beta.vectorbase.org/vectorbase.beta/app/>
 - b. Begin typing Protein motif in the Search For... filter and then choose Protein Motif Pattern.
 - c. At the search form, enter the protein regular expression, choose *Anopheles gambiae* str. PEST as the organism and click Get Answer.
Protein regular expression=
C.[DNEHQSTI]C.{4,6}[ST].{2}[WM][HR][RKENAMSLPGQT].{3,4}[GNEP].{3,6}C[NES][ASNR]C

Search for...

Genes

- Immunology
- Epitope Presence (IEDB)
- Protein features and properties
 - InterPro Domain
 - Isoelectric Point
 - Molecular Weight
- Protein targeting and localization
 - Predicted Signal Peptide
 - Transmembrane Domain Count
- Sequence analysis
 - Protein Motif Pattern

Identify Genes based on Protein Motif Pattern

Pattern

Organism

Note: You may select up to 5 values for this parameter.

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

- Arthropoda
 - Insecta
 - Diptera
 - Culicidae
 - Anopheles
 - Anopheles gambiae
 - Anopheles gambiae str. PEST

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

[Get Answer](#)

- d. Your results are 5 genes (7 transcripts) whose protein products contain the motif. While three of these genes are already annotated as GATA-binding proteins, two genes (3 transcripts) returned by the search are annotated as unspecified product. In one search we identified functions for two genes with previously unknown functions!

My Search Strategies

Opened (1) All (1) Public (4) Help

Unnamed Search Strategy *

Prot Motif 5 Genes

+ Add a step

Step 1

5 Genes (5 ortholog groups) [Revise this search](#)

Gene Results [Genome View](#) [Analyze Results](#)

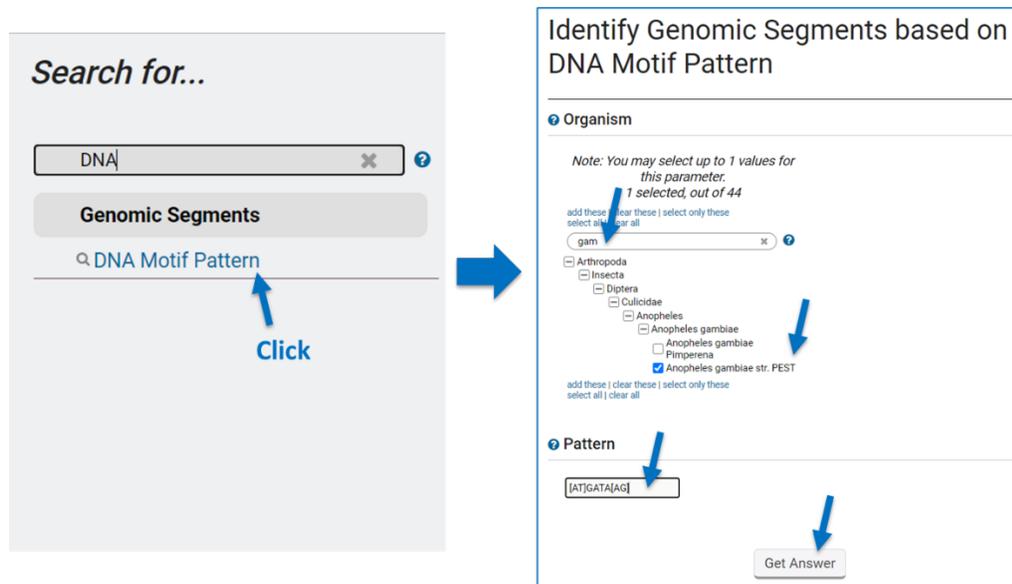
Genes: 5 Transcripts: 8 Show Only One Transcript Per Gene

Rows per page: 20 [Download](#) [Add to Basket](#) [Add Columns](#)

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Match Locations
AGAP002235	AGAP002235-RA	<i>Anopheles gambiae str. PEST</i>	AgamP4_2R:17,971,008..18,005,899(-)	GATA-binding protein 4/5/6 [Source:VB Community Annotation]	(120-144), (177-201)
AGAP002236	AGAP002236-RA	<i>Anopheles gambiae str. PEST</i>	AgamP4_2R:18,013,752..18,028,826(-)	unspecified product	(503-527), (566-590)
AGAP002236	AGAP002236-RB	<i>Anopheles gambiae str. PEST</i>	AgamP4_2R:18,013,752..18,028,826(-)	unspecified product	(573-597), (636-660)
AGAP002238	AGAP002238-RA	<i>Anopheles gambiae str. PEST</i>	AgamP4_2R:18,033,675..18,055,975(-)	unspecified product	(747-771), (801-825)
AGAP004228	AGAP004228-RA	<i>Anopheles gambiae str. PEST</i>	AgamP4_2R:52,246,051..52,346,042(-)	GATA-binding protein 1/2/3 [Source:VB Community Annotation]	(338-362), (397-421)
AGAP004228	AGAP004228-RB	<i>Anopheles gambiae str. PEST</i>	AgamP4_2R:52,246,051..52,346,042(-)	GATA-binding protein 1/2/3 [Source:VB Community Annotation]	(277-251), (286-310)

[COMMUNITY CHAT](#)

2. Start a new strategy by initiating a DNA Motif Pattern search.
 - a. Choose to search *Anopheles gambiae str. PEST*.
 - b. Enter the regular expression that represents the GATA site consensus sequence [AT]GATA[AG] into the **Pattern** field.
 - c. Click **Get Answer**



- d. Your results are 'segments' of DNA that contain the motif. Each segment has a known location within the *Anopheles gambiae* str. PEST genome. The search returned over 500,000 motifs!

The screenshot shows a search results page for '511,144 Genomic Segments'. At the top, there is a 'DNA Motif' section with '511,144 Segments' and an 'Add a step' button. Below this, a table displays the results. The table has columns for Segment ID, Organism, Genomic Location, and Motif. The first three rows are visible:

Segment ID	Organism	Genomic Location	Motif
AAAB01000047:10251-10257:r	Anopheles gambiae str. PEST	AAAB01000047:10,251..10,257 (-)	...TTTCTACGCGCATCTTTAATTGATAGATAATATGCTTGCCGCTGGT...
AAAB01000047:10989-10995:f	Anopheles gambiae str. PEST	AAAB01000047:10,989..10,995 (+)	...AAATAAATTGTAGTCTTCAAAGATAATGGTAACTTAACATCGATA...
AAAB01000047:12244-12250:f	Anopheles gambiae str. PEST	AAAB01000047:12,244..12,250 (+)	...CCACACCTCTCACAAAGGTGTGATAAATGTAGTTATAATTTTCCTA...

3. Colocate these GATA motifs with *Anopheles gambiae* str. PEST genes.
 - a. Click **Add a step** and choose **Use Genomic Colocation** to combine with other features.
 - b. Choose to run **A new search** for **Genes** based on **Taxonomy, Organism**

- c. At the search form choose *Anopheles gambiae* str. PEST for the **Organism** parameter.

- d. Arrange the colocation tool to read “Return each **gene from the new step** whose **upstream 200bp** region **overlaps** the **exact** region of the genomic segment from the current step and is on **either** strand”.

View | Analyze | Insert step before | Orthologs | Delete

Details for step **Genes by Rel Loc**
 4519 Genes

*Return each **Gene from Step 2** whose **upstream region** overlaps the **exact region** of a Genomic Segment from Step 1 and is on **either strand**

Exact
 Upstream: bp
 Downstream: bp
 Custom:
 begin at: - bp
 end at: - bp

Exact
 Upstream: bp
 Downstream: bp
 Custom:
 begin at: - bp
 end at: - bp

Revise

- e. The result is a list of genes whose upstream 200bp regions contain the GATA binding motif. The Region column shows the location of the motif.

Organism
13,796 Genes

DNA Motif
511,144 Segments

Step 1

Step 2
4,519 Genes

+ Add a step

4,519 Genes (3,610 ortholog groups)

Gene Results | Genome View | Analyze Results

Genes: 4,519 | Transcripts: 5,029 | Show Only One Transcript Per Gene

1 2 3 ... 252 | Rows per page: 20

Download | Add to Basket | Add C

Gene ID	Transcript ID	Organism	Genomic Location (Transcript)	Product Description	Match Count	Region
AGAP000422	AGAP000422-RA	<i>Anopheles gambiae</i> str. PEST	AgamP4_X:7593777..7598788(+)	(heparan sulfate)-glucosamine 3-sulfotransferase 3 [Source:VB Community Annotation]	1	7593577 - 7593776 (+)
AGAP010885	AGAP010885-RA	<i>Anopheles gambiae</i> str. PEST	AgamP4_3L:12478591..12480731(-)	(S)-2-hydroxy-acid oxidase [Source:VB Community Annotation]	2	12480761 - 12480960 (-)
AGAP005618	AGAP005618-RA	<i>Anopheles gambiae</i> str. PEST	AgamP4_2L:17920226..17920992(-)	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase [Source:UniProtKB/TrEMBLAcc:Q7Q6X6]	2	17921263 - 17921462 (-)
AGAP007113	AGAP007113-RA	<i>Anopheles gambiae</i> str. PEST	AgamP4_2L:42963004..42964452(+)	1-acyl-sn-glycerol-3-phosphate acyltransferase gamma [Source:VB Community Annotation]	1	42957794 - 42957993 (+)