


Genome Network Project

Cis-element analysis system

Cis-Finder

User Manual

contents

1.	About Cis-Finder	3
2.	Aanalysis Execution	4
2.1	Loading Query Sequence	4
2.2	Obtaining query sequence using BioMart	5
2.3	Parameter Settings	7
2.4	Parameter settings for each algorithm	8
2.5	Analysis execution 	9
3.	Analysis Result	10
3.1	Displaying analysis result	10
3.2	Motif search and clustering execution	11
3.3	Motif search and clustering result	13
3.4	Gene Ontology information	15
3.5	Gene List	16

1. About Cis-Finder

Cis-Finder is a tool that predicts candidate motifs (called Cis Elements candidates) from DNA sequences in genetic promoter region.

Cis-Finder will prevent omission on detecting motifs. Since it executes analyzing with combined representative algorithms for DNA sequence prediction, it achieves to detect characteristic patterns of motifs utilizing predicted data from those algorithms.

4 algorithms/tools (Consensus, MEME, Gibbs Sampler, MDSCan) are used in combination.

You can set parameters in each algorithms.

Followings are the main use of Cis-Finder.

1. To predict sequences that exist

- ① predict the sequence information that exists in common within the transcription control region of functionally relevant gene(transcript factor binding site).
- ② To predict common sequence present in the transcription control region of genes in different species evolved from a common ancestral gene (orthologous gene).

Figure1 Top Page of Cis-Finder



2. Aanalysis Execution

2.1 Loading query sequence

For analysis, you are expected to prepare query sequence data in multi-FASTA format. To load data, select one of those procedures below.

①copy & paste ...

input the query sequence by copy-and-paste.

②file upload ...

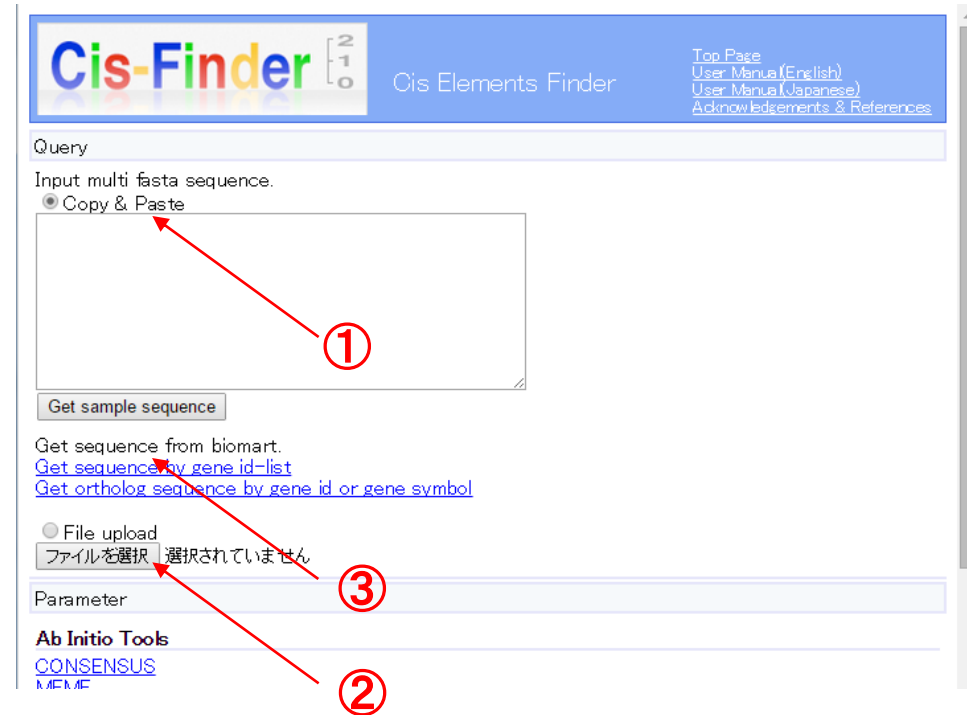
upload data file from a local disk.

③Biomart ...

Obtain the query sequence from Biomart.

(reference[2.2 Obtaining query sequence using BioMart.])

Figure 2 Query sequence specified Cis-Finder



2.2 Obtaining query sequence using BioMart.

Cis-Finder holds search function of BioMart and is able to get ortholog gene sequences of human and other species registered in the Ensembl database.

(1) BioMart keyword search & species specification. (Figure3)

When you click the link as ③ of Figure2 points, BioMart Keyword search screen appears (Figure3)。

① Keyword ...

- If you want to search by keywords, select the “Gene Symbol” radio button, and enter the keyword (Keyword is expected to be the exact match).
- If you want to search by the Ensembl ID, select the “Ensembl Gene ID” radio button, and then enter the ID. (The ID is expected to be the human Ensembl Gene ID)

② Ortholog ... specify the species.

③ When you press the “Get Gene ID-List” button, gene & region setting screen shows up (Figure4).

(2) BioMart gene search: gene & region settings (Figure4)

① Sequence Type ... specify the type of sequence.

② Sequence Region ... If you choose “_flank” at the Sequence Type, select upstream or downstream of sequence and specify the search range in number.

③ When you press the “Get Sequence” button, you will get sequence with customized condition you made.

Figure 3 BioMart keyword search & species specification.

Get Sequence From Biomart
Get Ortholog Sequence By Gene ID or Gene Symbol

Sequence Dataset: Homo sapiens (human)

Keyword:

Orthologs:

- Gene Symbol
- Gene ID
- Ensembl Gene ID(s):
- Homo sapiens (human)
- Anopheles gambiae (mosquito)
- Bos taurus (cow)
- Caenorhabditis elegans (nematode)
- Canis familiaris (dog)
- Ciona intestinalis (transparent sea squirt)
- Ciona savignyi (Pacific transparent sea squirt)
- Danio rerio (zebrafish)
- Drosophila melanogaster (fruit fly)
- Gallus gallus (chicken)
- Macaca mulatta (rhesus macaque)
- Monodelphis domestica (opossum)
- Mus musculus (house mouse)
- Pan troglodytes (chimpanzee)
- Rattus norvegicus (Norway rat)
- Saccharomyces cerevisiae (baker's yeast)
- Takifugu rubripes (torafugu)
- Tetraodon nigroviridis (spotted green pufferfish)
- Xenopus tropicalis (western clawed frog)

Get Gene ID-List

Copyright © 2008–2015 National Institute of Genetics

Figure 4 BioMart gene search: gene & region settings

Get Ortholog Sequence By Gene ID or Gene Symbol

Sequence Dataset	Homo sapiens (human)	
Gene ID	Homo sapiens (human)	ENSG00000141510
Ortholog Gene ID-List	Mus musculus (house mouse)	ENSMUSG00000059552
	Pan troglodytes (chimpanzee)	ENSPTRG00000008703
	Rattus norvegicus (Norway rat)	ENSRNOG00000046857

Sequence Type:

- Unspliced (Transcript)
- Unspliced (Gene)
- Flank (Transcript)
- Flank (Gene)
- Flank-coding region (Transcript)
- Flank-coding region (Gene)
- 5' UTR
- 3' UTR
- Exon sequences
- cDNA sequences
- Coding sequence
- Protein

Sequence Region:

Upstream Downstream

1000

Get Sequence

Copyright © 2008–2015 National Institute of Genetics

(3) Obtaining sequence data (Figure 5)

- When the sequence data successfully being acquired, it appears in the “Results” area as Figure 5 above.
- When you click the “Paste Sequence” button, the sequence data which shows in the results area will be copied and pasted to the query pane as Figure 5 below indicates.

Figure 5 Obtaining sequence data

Sequence Type

- Unspliced (Transcript)
- Unspliced (Gene)
- Flank (Transcript)
- Flank (Gene)
- Flank-coding region (Transcript)
- Flank-coding region (Gene)

Sequence Region

1000

Get Sequence

Results

```
>ENS00000141510
T CCTCTTCTGGGAGT A GGCAGAGA CT CCGGGAGGAGAGGCGAA CAGCGGA
CGCCAAATCTTTTGAAGCACTGTGTT CCTTAGCACCGGGGTGCTACGGG
CCTCTTGCTGTCGGGGATTTGGTCCACCTCCGAT TGGCCCGCCGATCC
CGGATCAGATTTCCGGGGGACCCACGGAACCGCGGAGCCGGGACGTGAAA
GGT TAGAAGGTTTCCCGT TCCCATCAAGCCCTAGGGCTCCTGTTGGCTGCTG
GGAGTTGTAGTCTGAA CGCTTCTATCTTGGCGA GAA GGCCTACGCTCCCTAC
TACCGAGTCCCGGGTAA TTTCTTAAAGCACCTGCACCGCCCGCCCGCCGCTGCA
GAGGGGCGACAGGTCTTGCACCTCTTCTGCATCTCATTCTCCAGGCTT
CAGACCTGTCTCCCTCATTCAAAAAATTTATTATCGAGCTCTTACTTGGT
```

Paste Sequence

Copyright © 2008–2015 National Institute of Genetics

Cis-Finder Cis Elements Finder

Top Page
User Manual (English)
User Manual (Japanese)
Acknowledgements & References

Query

Input multi fasta sequence.

- Copy & Paste

```
>ENS00000141510
AGTTCTCAGGGATCCGACGCA GAGCTAAAGAAA CCGACCTGTGCTTCCCTCC
TCTTCTGGGAGT A GGCAGAGA CT CCGGGAGGAGAGGCGAA CAGCGSA GSC
CAATCTTTTGAAGCACTGTGTT CCTTAGCACCGGGGTGCTACGGGCT
CTTGCTGTCGGGGATTTGGTCCACCTCCGAT TGGCCCGCCGATCCCGG
ATCAGATTTCCGGGGGACCCACGGAACCGCGGAGCCGGGACGTGAAAAGST
TAGAAGGTTTCCCGT TCCCATCAAGCCCTAGGGCTCCTGTTGGCTGCTGCGGA
GTTGTAGTCTGAA CGCTTCTATCTTGGCGA GAA GGCCTACGCTCCCTAC
CGAGTCCCGGGTAA TTTCTTAAAGCACCTGCACCGCCCGCCCGCCGCTGCA
GAGGGGCGACAGGTCTTGCACCTCTTCTGCATCTCATTCTCCAGGCTTCA G
```

Get sample sequence

Get sequence from biomaart.
[Get sequence by gene id-list](#)
[Get ortholog sequence by gene id or gene symbol](#)

File upload
ファイルを選択 | 選択されていません

Parameter

Ab Initio Tools
[CONSENSUS](#)
[MEME](#)

2.3 Parameter Settings

As shown in Figure 6, you can specify various parameters for the cis-element analysis.

① Parameter settings for the representative algorithms.

You can specify parameters for each algorithm of Consensus, MEME, Gibbs sampler, MDScan and TFSCAN. (There is no parameter to TRANSFAC)

Figure 6 Parameter Settings

The screenshot shows a web interface for parameter settings. At the top, there is a text area containing a DNA sequence: TACGGGCTCTTGCTGTCGCGGGATTTCGGTCCACCTTCGATTTGGCCGCCGCATCCCGGATCAGATTCGCGGGGACCCACGGAAACCCGCGGAGCCGGACGTGAAAGGTTAGAAGGTTTCCCGTTCATCAAGCCCTAGGCTCCTCGTGGCTGCTGGGAGTTGTAGTCTGAACGCTTCTATCTTGGCAGAAGCGGCTACGCTCCCGCTACCGAGTCCCGCGGTAATCTTAAAGCACC TGACCCGCCCCCGCGCTGCGAGGGGCGCAGAGTCTTGCACCTC. Below the sequence, there are links: "Get sequence from biomaRT", "Get sequence by gene id-list", and "Get ortholog sequence by gene id or gene symbol". There is a "File upload" section with a text input field and a "浏览..." button. The "Parameter" section is titled "Ab Initio Tools" and lists several algorithms with links: CONSENSUS, MEME, GIBBS, MDSCAN, TFSCAN, TFSCAN, TRANSFAC, and TRANSFAC. There is an "Execute" button at the bottom of this section. The footer of the page says "2008 Genome Network Project". A red circle with the number 1 is positioned to the left of a red bracket that groups the "Ab Initio Tools" section.

2.4 Parameter settings for each algorithm

The parameter setting screen of each algorithm is shown in Figures 7 to 10. To return to the default value, click the “set default” button at the top of a setting table.

Figure 7 Parameter settings of Consensus

The screenshot shows the 'Consensus' parameter settings. At the top, there is a 'set default' button. Below it, a table lists various parameters with their current values and input fields. The parameters include search motif length, matrix saving, cycle requirements, minimum distance, and options for handling nucleic acid sequences. A 'set default' button is also present at the top right of the table.

Parameter	Value
Length of search motif	8
Number of matrix to save	1000
Number of cycles if 0 or more motifs per sequence are required	10
Number of cycles if 1 or more motifs per sequence are required	
Minimum distance between words	8
Terminate indicated number of cycles after most significant alignment	
Use designated prior frequencies	<input type="checkbox"/>
Seed with first sequence and proceed linearly through list	<input type="checkbox"/>
Options for handling the complement of nucleic acid sequences the four options in this section are mutually exclusive	Include both strands as a single sequence (i.e., orientation unknown)
SelectOne	Save the top progeny for each parental matrix
Number of top matrices to print	4
Number of final matrices to print	2

Figure 8 Parameter settings of MEME

The screenshot shows the 'MEME' parameter settings. It features a 'set default' button at the top. The settings table includes parameters for search motif length (upper and lower limits), number of sites, maximum number of motifs, motif distribution menu, and various alignment options. A 'set default' button is located at the top right of the table.

Parameter	Value
MAXimum length of search motif (upper limit 300)	10
MINimum length of search motif (lower limit 2)	6
MAXimum number of sites for each motif (upper limit 300)	
MINimum number of sites for each motif (lower limit 2)	
MAXimum number of motifs to find	
Motif distribution menu	Any number of repetitions per sequence
Stop if E-motif value is greater than	
Maximum EM iterations to run	50
Do not adjust motif length using multiple alignment	<input checked="" type="checkbox"/>
Gap opening cost for multiple alignments	11
Gap extension cost for multiple alignments	1
Do not count end gaps in multiple alignments	<input checked="" type="checkbox"/>
Use complementary strand	<input checked="" type="checkbox"/>
Force palindromes	<input type="checkbox"/>

Figure 9 Parameter settings of Gibbs sampler

The screenshot shows the 'GIBBS' parameter settings. It includes a 'set default' button at the top. The settings table covers parameters such as search motif length, expected number of elements, near optimal cutoff, random number generator seed, sampling runs, iterations between local maxima, iterations per seed, pseudo count weights, and site weights. A 'set default' button is at the top right of the table.

Parameter	Value
Length of search motif	8
Expected number of elements for each type	10
Near optimal cutoff (%)	50
Give seed for random number generator	1000
Maximum number of sampling runs	10
Number of iterations between successive local maxima	20
Maximum number of iterations for each seed	500
Pseudo count weights	0.1
Pseudo site weights (sites sampler)	0.8
DO NOT use fragmentation (i.e., column sampler)	<input type="checkbox"/>
Use element order in probabilities (sites sampler)	<input type="checkbox"/>
Randomly shuffle input sequences	<input type="checkbox"/>
DON'T remove protein low complexity regions	<input type="checkbox"/>

Figure 10 Parameter settings of MDScan

The screenshot shows the 'MDSCAN' parameter settings. It features a 'set default' button at the top. The settings table includes parameters for motif width, number of top sequences for candidate motifs, expected bases per motif site, background distribution file, background sequence file, number of candidate motifs to scan and refine, number of top motifs to report, and number of refinement iterations. A 'set default' button is at the top right of the table.

Parameter	Value
motif width	8
number of top sequences to look for candidate motifs	5
number of top sequences to confirm candidate motifs	
expectd bases per motif site in the top sequences	
background distribution file	Query Sequence
background sequence file	input sequences
number of candidate motifs to scan and refine	30
number of top motifs to report at the end	2
number of refinement iterations	10

2.5 Analysis execution

① presses the “Execute” button to start the analysis.

Figure 11 Analysis execution

The screenshot shows a web interface for sequence analysis. At the top, there is a text area containing a DNA sequence: TACGGGCTCTTGCTGTCGCGGGATTTCGGTCCACCTTCGATGGGCCGCCGATCCCGGATCAGATTTCGCGGGGACCCACGGAAACCCGCGGAGCCGGACGTGAAAGGTTAGAAGGTTTCCCGTTCCCATCAAGCCCTAGGGCTCCTCGTGGTCTGGGAGTTGTAGTCTGAACGCTTCTATCTTGGCGAGAAGCGGCTACGCTCCCGCTACCGAGTCCCGCGGTAATCTTAAAGCACC TGACCCGCCCCCGCCGCTGCGAGGGGCGCAGCAGGTCTTGCACCTC. Below the text area, there are links: "Get sequence from biomaRT", "Get sequence by gene id-list", and "Get ortholog sequence by gene id or gene symbol". There is a "File upload" section with a text input field and a "参照..." button. A "Parameter" section is visible. Under "Ab Initio Tools", there are links for CONSENSUS, MEME, GIBBS, MDSCAN, TFSCAN, and TRANSFAC. The "Execute" button is highlighted with a red circle and the number 1, with an arrow pointing to it. The footer of the page reads "2008 Genome Network Project".

3. Analysis Result

3.1. Displaying analysis result

After the analysis being completed, you can view detailed result for each sequence as shown in Figure 12.

Figure 12 Displaying analysis result



3.2 Motif search and clustering execution.

If you click on the colored box (①) which implies a motif, its sequence and location will be shown on the left side of the window (②).

图13 selection of motif

The screenshot displays a software interface for motif search and clustering. It is divided into two main panels: 'Motif Information' on the left and 'Motif View' on the right.

Motif Information Panel:

- Ab Initio Tools:** Contains a 'Motif ID' and 'Motif Logo' section. It lists three motifs: 'consensus1' (TGCCCTCA), 'consensus2' (CAGGATC), and 'meme1' (TTTCCCTCA), each with a corresponding logo.
- Motif Search & Clustering:** Contains a 'Selected Motifs' table and a 'Choose Search Set' section.

Order No	Motif Sequence(*)
1	TGATAAG
2	CAGGAKTC
3	TGCCCTYA
4	CTGACTCT

Motif View Panel:

- Shows a 'Query1: ENSG00000141510' sequence.
- Includes a 'CONSENSUS' line and a 'MEME' logo.
- Displays a grid of motifs (tfscan1 to tfscan17) with colored boxes (blue, green, yellow) indicating motif locations. A red circle (①) highlights a blue box in the top row, with a red arrow pointing to it.

A red circle (②) is placed on the 'Selected Motifs' table in the left panel, with a bracket indicating that clicking on a colored box in the 'Motif View' panel (①) updates this table.

☒14 search motif and execute clustering

① Selection of Database ...

Select the database you want to search with.

•Ensembl ... Sequences upstream or downstream 1kb of Ensembl genes of human and mouse

•CAGE ... sequence of unknown transcription start point

② To start searching, please click the "Search & Clustering" button.

④ To clear motif selections, please press "Clear" button.

Motif Information

Ab Initio Tools

Motif ID Motif Logo

consensus1 TGCCCTCA

consensus2 CAGGATC

meme1 TTTCCCTCA

Order No	Motif Sequence(*)
1	TGATAAG
2	CAGGAKTC
3	TGCCCTYA
4	CTGACTCT

Choose Search Set

Query

Database

Human -> Ensembl 1000bp Upstream

Search & Clustering Clear

Motif View

Query1:ENSG00000141510

CONSENSUS

MEME

SITES

MOTIF

tfscan1

tfscan2

tfscan3

tfscan4

tfscan5

tfscan6

tfscan7

tfscan8

tfscan9

tfscan10

tfscan11

tfscan12

tfscan13

tfscan14

tfscan15

tfscan16

tfscan17

Figure 15 motif search and clustering result

3.3 Motif search and clustering result.

When the motif search being completed, you will see the screen shown in Figure 15.

① Selected Motifs ...

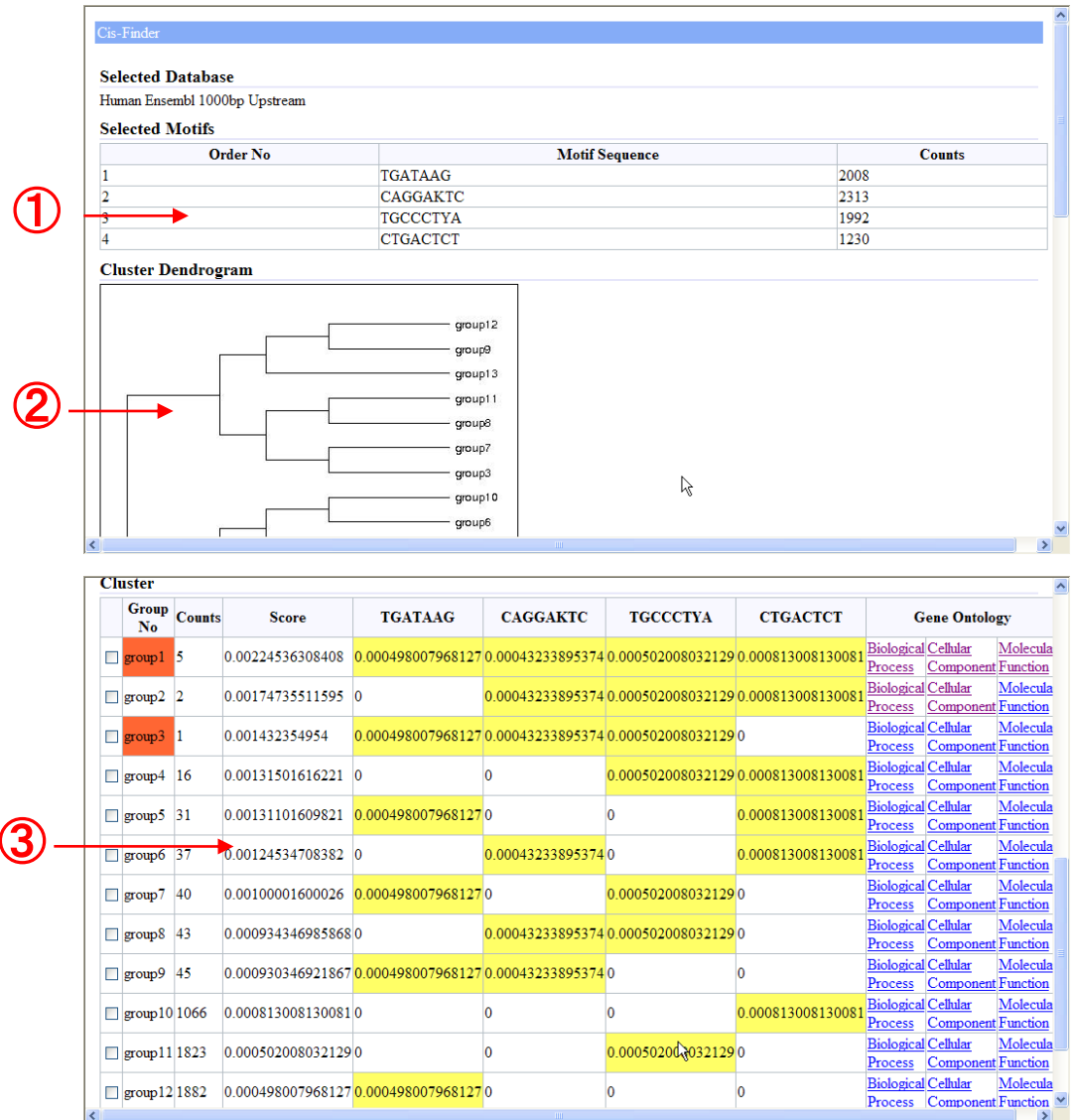
Displays sequence pattern and number of hits of the selected motif.

② Cluster Dendrogram ...

shows the clustering results.

③ Cluster ...

Genes from the specified database are grouped by whether each of those has a target motif or not. The Score row applies to inverse value of the number of hits.



☒16 detail of each cluster

You can confirm detailed information for clusters by gene list and Gene Ontology information.

①Gene Ontology***

You can view Gene Ontology information belongs to each cluster.

→3.4 For details, please see 3.4 Gene Ontology information.

②Gene List***

Please mark a checkbox for the groups (②) that you want to view, and click the "Gene List" button (③).

→3.5 For details, please see 3.5 Gene List.

②

<input type="checkbox"/>	group1	5	0.00224536308408	0.000498007968127	0.00043233895374	0.000502008032129	0.000813008130081	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group2	2	0.00174735511595	0	0.00043233895374	0.000502008032129	0.000813008130081	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group3	1	0.001432354954	0.000498007968127	0.00043233895374	0.000502008032129	0	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group4	16	0.00131501616221	0	0	0.000502008032129	0.000813008130081	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group5	31	0.00131101609821	0.000498007968127	0	0	0.000813008130081	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group6	37	0.00124534708382	0	0.00043233895374	0	0.000813008130081	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group7	40	0.00100001600026	0.000498007968127	0	0.000502008032129	0	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group8	43	0.000934346985868	0	0.00043233895374	0.000502008032129	0	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group9	45	0.000930346921867	0.000498007968127	0.00043233895374	0	0	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group10	1066	0.000813008130081	0	0	0	0.000813008130081	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group11	1823	0.000502008032129	0	0	0.000502008032129	0	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group12	1882	0.000498007968127	0.000498007968127	0	0	0	Biological Cellular Process	Component	Molecular Function
<input type="checkbox"/>	group13	2140	0.00043233895374	0	0.00043233895374	0	0	Biological Cellular Process	Component	Molecular Function

Gene List

2008 Genome Network Project

③

①

Figure 16 Gene Ontology information

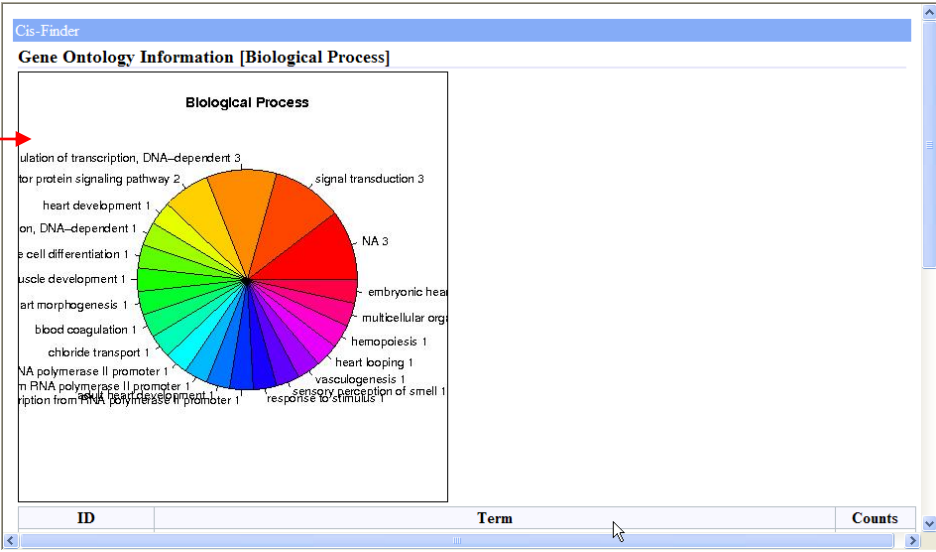
3.4 Gene Ontology Information

You can browse following Gene Ontology information in this pane.

- Biological Process
- Cellular Component
- Molecular Function

① Display a pie chart which indicates ratio of Gene Ontology Term for genes belong to the cluster.

①



② Display a list of Gene Ontology Term.

②

ID	Term	Counts
NA	NA	3
GO:0007165	signal transduction	3
GO:0006355	regulation of transcription, DNA-dependent	3
GO:0007186	G-protein coupled receptor protein signaling pathway	2
GO:0007507	heart development	1
GO:0045893	positive regulation of transcription, DNA-dependent	1
GO:0055007	cardiac muscle cell differentiation	1
GO:0048738	cardiac muscle development	1
GO:0003007	heart morphogenesis	1
GO:0007596	blood coagulation	1
GO:0006821	chloride transport	1
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1
GO:0006357	regulation of transcription from RNA polymerase II promoter	1
GO:0007512	adult heart development	1
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	1
GO:0050896	response to stimulus	1
GO:0007608	sensory perception of smell	1
GO:0001570	vasculogenesis	1
GO:0001947	heart looping	1
GO:0030097	hemopoiesis	1
GO:0007275	multicellular organismal development	1

3.5 Gene List

Display gene list belongs to the cluster. In addition, it shows information of gene expression and mutual interaction between proteins. (Link to genome network platform)

① Display information of gene annotation.

② Heatmap Viewer ... display expression information for each tissues. (Figure 18)

③ PPI Network Viewer ... Display information of Protein-protein Interactions. (Figure 19)

② ③ Figure 17 Gene List

No	Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	Gene Name	Chromosome	Start	End	Description
1	ENSG00000092203	ENST00000262709	9878	TOX4	14	21015175	21037155	TOX high mobility group box family member 4 (Epidermal Langerhans cell protein LCP1). [Source:Uniprot/SWISSPROT,Acc:O94842]
2	ENSG00000105650	ENST00000262805	5143	PDE4C	19	18179771	18198225	"cAMP-specific 3',5'-cyclic phosphodiesterase 4C (EC 3.1.4.17) (DPDE1) (PDE21). [Source:Uniprot/SWISSPROT,Acc:Q08493]"
3	ENSG00000122783	ENST00000361743						
4	ENSG00000137968	ENST00000370856		SLC44A5	1	75441785	75849368	Choline transporter-like protein 5 (Solute carrier family 44 member 5). [Source:Uniprot/SWISSPROT,Acc:Q8NCS7]
5	ENSG00000165204	ENST00000277309	392392	OR1K1	9	124602223	124603173	Olfactory receptor 1K1. [Source:Uniprot/SWISSPROT,Acc:Q8NGR3]
6	ENSG00000168288	ENST00000375738		C2orf25	2	150134399	150147914	"Uncharacterized protein C2orf25, mitochondrial precursor. [Source:Uniprot/SWISSPROT,Acc:Q9H3L0]"
7	ENSG00000183072	ENST00000329198	1482	NKX2-5	5	172591744	172594868	Homeobox protein Nkx-2.5 (Homeobox protein NK-2 homolog E) (Cardiac-specific homeobox) (Homeobox protein CSX). [Source:Uniprot/SWISSPROT,Acc:P52952]
8	ENSG00000184908	ENST00000375667		CLCNKB	1	16247871	16256390	Chloride channel protein ClC-Kb (Chloride channel Kb) (ClC-K2).

Figure 18 Heatmap Viewer

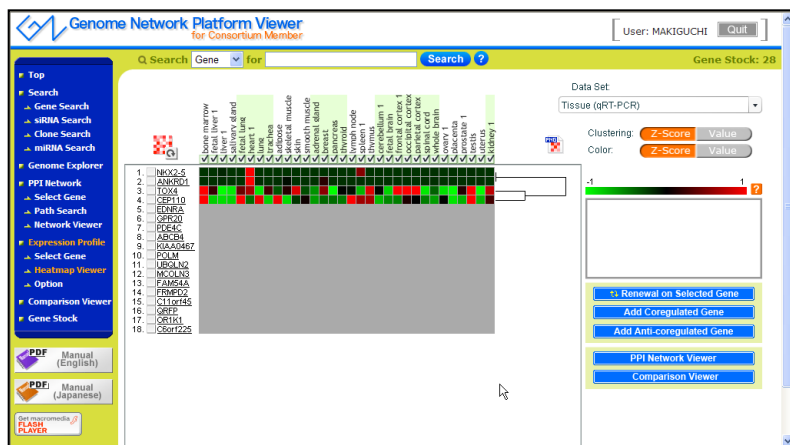


Figure 19 Information of protein-protein interactions

