

Download information of all IMG genes annotated with a specific function

(6/25/2020)

This document describes how IMG users can download information of all IMG genes annotated with a specific function.

Let's assume that the function we are interested in is enzyme **EC:3.2.1.14** Chitinase. Also, we will only use isolate genomes in this example. Users can follow similar steps to find metagenome genes.

Warning: IMG is a shared resource. So PLEASE do NOT try to download all IMG genes with all functions!!! Yours jobs will be terminated if you try to do that.

Step 1: Find all genomes having genes annotated with EC:3.2.1.14

From IMG UI, go to **Find Functions -> Enzyme**, and type "3.2.1.14" and click the **Apply** button.

[Add Selected to Function Cart](#)
[Select All](#)
[Clear All](#)

Filter column: Filter : [Apply](#) [?](#)

[Export](#) Page 1 of 1 << first < prev **1** next > last >>

[Column Selector](#) [Select Page](#) [Deselect Page](#)

Select	ID	Name	Isolate Genome Count	Metagenome Count
<input type="checkbox"/>	EC:3.2.1.14	Chitinase.	24353	26851
<input type="checkbox"/>	EC:3.2.1.140	Lacto-N-biosidase.	0	0
<input type="checkbox"/>	EC:3.2.1.141	4-alpha-D-((1->4)-alpha-D-glucano)trehalose trehalohydrolase.	16959	19773
<input type="checkbox"/>	EC:3.2.1.142	Limit dextrinase.	0	0
<input type="checkbox"/>	EC:3.2.1.143	Poly(ADP-ribose) glycohydrolase.	405	10803
<input type="checkbox"/>	EC:3.2.1.144	3-deoxyoctulosonase.	0	0
<input type="checkbox"/>	EC:3.2.1.145	Galactan 1,3-beta-galactosidase.	0	0
<input type="checkbox"/>	EC:3.2.1.146	Beta-galactofuranosidase.	0	0
<input type="checkbox"/>	EC:3.2.1.147	Thioglucosidase.	11	205
<input type="checkbox"/>	EC:3.2.1.148	S-ribosylhomocysteine lyase.	0	0
<input type="checkbox"/>	EC:3.2.1.149	Beta-primeverosidase.	0	0

[Export](#) Page 1 of 1 << first < prev **1** next > last >>

[Add Selected to Function Cart](#)
[Select All](#)
[Clear All](#)

Click the Isolate Genome Count number to open the genome list in a separate window. (Please note that you may see different counts because the number of IMG genomes can change daily.)

Isolate Genomes with EC:3.2.1.14

Genomes with *Chitinase*.

hint: The count(s) between gene and function are unique hits. In another word, if a gene hits the same function in multiple times at different positions, it's counted only once.

*Showing counts for all applicable genomes.

[Show only counts via Genome Cart](#)

Domains(D): * = Microbiome,

B = Bacteria, A = Archaea, E = Eukarya, P = Plasmids, G = GFragment, V = Viruses.

Genome Completion(C): F = Finished, P = Permanent Draft, D = Draft.

Add Selected to Genome Cart

Select All

Clear All

24353 of 24353 rows selected

Filter column: Domain Filter text Apply

Export

Page 1 of 244

<< first < prev

1

2

3

4

5

6

7

8

9

10

next >

last >>

100

▼

Column Selector

Select Page

Deselect Page

Select	Domain	Status	Genome	Gene Count
<input checked="" type="checkbox"/>	A	P	Haloarchaeobius iranensis IBRC-M 10013	1
<input checked="" type="checkbox"/>	A	F	Thermococcus chitonophagus GC74	2
<input checked="" type="checkbox"/>	A	P	Natrialba asiatica DSM 12278	2
<input checked="" type="checkbox"/>	A	F	Haloferax mediterranei R-4	1
<input checked="" type="checkbox"/>	A	P	Haloferax larsenii CDM_5	1

Select all genomes in the list by clicking the **Select All** button. Then scroll down to save the selected genomes into a Workspace genome set called **ec_3_2_1_14**:

<input checked="" type="checkbox"/>	B	P	Fangia hongkongiensis FSC776	2
<input checked="" type="checkbox"/>	B	P	Sinorhizobium arboris LMG 14919	1
<input checked="" type="checkbox"/>	B	P	Photobacterium angustum A2-4	5
<input checked="" type="checkbox"/>	B	P	Klebsiella pneumoniae S_19PV	1
<input checked="" type="checkbox"/>	B	P	Parabacteroides sp. AF27-14	1

Export

Page 1 of 244

<< first < prev

1

2

3

4

5

6

7

8

9

10

next >

last >>

100

▼

24353 of 24353 rows selected

Add Selected to Genome Cart

Select All

Clear All

Save Genomes to My Workspace

hint: Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

Save **selected genomes** to [My Workspace](#).

(Special characters in file name will be removed and spaces converted to _)

Save to File name:

Append to the following genome set:

Replacing the following genome set:

▼

Save Selected to Workspace

Click the **Save Selected to Workspace** button to save the data.

Now return to the **Enzymes** page to select *EC:3.2.1.14* to save it into a Workspace function set also called **ec_3_2_1_14**. (You are free to choose a different name as long as you remember what it is.)

Select	ID	Name	Isolate Genome Count	Metagenome Count
<input checked="" type="checkbox"/>	EC:3.2.1.14	Chitinase.	24353	26851
<input type="checkbox"/>	EC:3.2.1.140	Lacto-N-biosidase.	0	0
<input type="checkbox"/>	EC:3.2.1.141	4-alpha-D-((1->4)-alpha-D-glucano)trehalose trehalohydrolase.	16959	19773
<input type="checkbox"/>	EC:3.2.1.142	Limit dextrinase.	0	0
<input type="checkbox"/>	EC:3.2.1.143	Poly(ADP-ribose) glycohydrolase.	405	10803
<input type="checkbox"/>	EC:3.2.1.144	3-deoxyoctulosonase.	0	0
<input type="checkbox"/>	EC:3.2.1.145	Galactan 1,3-beta-galactosidase.	0	0
<input type="checkbox"/>	EC:3.2.1.146	Beta-galactofuranosidase.	0	0
<input type="checkbox"/>	EC:3.2.1.147	Thioglucosidase.	11	205
<input type="checkbox"/>	EC:3.2.1.148	S-ribosylhomocysteine lyase.	0	0
<input type="checkbox"/>	EC:3.2.1.149	Beta-primeverosidase.	0	0

Export Page 1 of 1 << first < prev 1 next > last >> All ▾

1 of 11 rows selected

Add Selected to Function Cart Select All Clear All

Save Functions to My Workspace

hint: Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

Save **selected functions** to [My Workspace](#).
(Special characters in file name will be removed and spaces converted to _)

Save to File name:

Append to the following function set.

Replacing the following function set:

Save Selected to Workspace

Step 2: Find all genes annotated with this function

Now go to **Workspace** -> **Genome Sets** to select the set **ec_3_2_1_14** you just created.

Genome Workspace Tips ?

Genome Cart

You have [29 Genome\(s\)](#) in your cart.

Genome Sets

Group sharing is not displayed. Please go to [Preferences](#) to change the sharing display options.

Genome Sets | Import & Export | Function Profile | Blast | Set Creation | Set Operation | Statistical Analysis NEW

1 of 4 rows selected

Filter column: File Name Filter text: Apply ?

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	File Name	Number of Genomes (click the link to each individual set)	File Size
<input type="checkbox"/>	cenocepacia	29	314 B
<input checked="" type="checkbox"/>	ec_3_2_1_14	24353	260.844 KB
<input type="checkbox"/>	no_genbank_id	1435	15.425 KB
<input type="checkbox"/>	no_portal	228	2.460 KB

Export Page 1 of 1 << first < prev 1 next > last >> All

1 of 4 rows selected

Add Selected to Public | Add Selected to Genome Cart | Select All | Clear All | Remove Selected

Then click the **Function Profile** tab to select the function set **ec_3_2_1_14**. Scroll down to find **Submission as Computation Job Using Message System**. Name the result file as **job_3_2_1_14** (or any other name you like), and click the **Submit Computation** button.

Genome Sets

Group sharing is not displayed. Please go to [Preferences](#) to change the sharing display options.

Genome Sets

Import & Export

Function Profile

Blast

Set Creation

Set Operation

Statistical Analysis

NEW

Genome Set Function Profile

hint: Due to computation and browser display limitations, the number of cells in the result table is restricted to 100,000, which is computed as the number of functions multiplied by the total number of genome sets + individual genomes in all sets.
It's easier to download your large dataset via [JGI Portal](#) through [Import & Export](#) tab. All data set downloads include all of IMG's gene functional annotations.

Use only functions in set:

Use all functions of type:

MER-FS Metagenome:

Display: counts per set counts per genome both

Genome Set Function Profile

Submit as Computation Job Using Message System

You may submit a genome set function profile computation to run in the background.

Save as a new job with name:

Replace the selected job:

Submit Computation

You will get a confirmation that the job has been successfully submitted. You can go to **Workspace -> My jobs** to check the status of that job. You will also get a notification when the job is completed.

Computation Jobs

All submitted jobs may take at least one day to complete.
You have 0 jobs.

1 of 2 rows selected

Filter column: Name Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Select	Name	Type	Start Time	Parameters	File Size	End Time	Status
<input type="checkbox"/>	cenocepacia_analysis	analysisStats	2020/06/24 10:47:13	pfam geneCount relative default	16.93 MB	2020/06/24 18:03:20	completed
<input checked="" type="checkbox"/>	job_3_2_1_14	Genome Function Profile	2020/06/22 15:27:37	--function ec_3_2_1_14 --genome ec_3_2_1_14 - -datatype assembled	4.84 MB	2020/06/22 18:20:29	completed

Export Page 1 of 1 << first < prev 1 next > last >> All

1 of 2 rows selected

Delete Download

When the job is completed, you can click the completed link to check the result. However, in many cases the result will be too huge, and the web browser will time out. An alternative way is to simply download the result – i.e., select the job and click the **Download** button.

The download result is a zip file. When you unzip the file, you will see:

C:\IMG\Data\job_3_2_1_14-jun-2020.zip\

Name	Size	Packed Size	Modified	Created	Accessed	Attributes	Encrypted	Comment	CRC	Method
done.txt	20	22	2020-06-22 18:20			-rw-rw-r--	-		CDCF3F25	Deflate
email.done.txt	25	24	2020-06-22 18:20			-rw-rw-r--	-		45A95A55	Deflate
info.txt	109	87	2020-06-22 15:27			-rw-rw-r--	-		63E5E73A	Deflate
list.txt	3 137 870	462 610	2020-06-22 18:20			-rw-rw-r--	-		938085FD	Deflate
log.txt	129	93	2020-06-22 18:20			-rw-rw-r--	-		4DCC42DD	Deflate
profile.txt	608 318	123 666	2020-06-22 18:20			-rw-rw-r--	-		F95D7B30	Deflate
set.txt	534 215	158 382	2020-06-22 18:20			-rw-rw-r--	-		4C2E2284	Deflate

1 / 7 object(s) selected 3 137 870 3 137 870 2020-06-22 18:20:28

The file “list.txt” contains the complete gene list. The 3rd column is the IMG gene OID, the 4th column is the gene product name, and the 5th column is the correspond IMG genome OID.

```

list - Notepad
File Edit Format View Help
ec_3_2_1_14 EC:3.2.1.14 637099575 chitinase B 637000277
ec_3_2_1_14 EC:3.2.1.14 637163698 chitinase domain protein 637000086
ec_3_2_1_14 EC:3.2.1.14 637165689 chitinase family 18 (EC:3.2.1.14) 637000154
ec_3_2_1_14 EC:3.2.1.14 637205987 chitinase family 18 (EC:3.2.1.14) 637000304
ec_3_2_1_14 EC:3.2.1.14 637268823 hydrolase 637000305
ec_3_2_1_14 EC:3.2.1.14 637269652 chitinase family 18 (EC:3.2.1.14) 637000305
ec_3_2_1_14 EC:3.2.1.14 637400993 chitinase family 18 (EC:3.2.1.14) 637000335
ec_3_2_1_14 EC:3.2.1.14 637455106 chitinase family 18 (EC:3.2.1.14) 637000074
ec_3_2_1_14 EC:3.2.1.14 637503596 chitinase family 18 (EC:3.2.1.14) 637000015
ec_3_2_1_14 EC:3.2.1.14 637790575 chitinase family 18 (EC:3.2.1.14) 637000013
ec_3_2_1_14 EC:3.2.1.14 637790800 alpha-amylase/chitinase/chitin-binding protein 637000013
ec_3_2_1_14 EC:3.2.1.14 637831538 chitinase family 18 (EC:3.2.1.14) 637000128
ec_3_2_1_14 EC:3.2.1.14 637842071 chitinase family 18 (EC:3.2.1.14) 637000052
ec_3_2_1_14 EC:3.2.1.14 637907992 chitinase family 18 (EC:3.2.1.14) 637000111
ec_3_2_1_14 EC:3.2.1.14 638001572 chitinase family 18 (EC:3.2.1.14) 637000219
ec_3_2_1_14 EC:3.2.1.14 638024292 chitinase family 18 (EC:3.2.1.14) 637000186
ec_3_2_1_14 EC:3.2.1.14 638032487 Chitinase 637000171
ec_3_2_1_14 EC:3.2.1.14 638119440 chitinase family 18 (EC:3.2.1.14) 637000259
ec_3_2_1_14 EC:3.2.1.14 638147373 chitinase family 18 (EC:3.2.1.14) 637000112
ec_3_2_1_14 EC:3.2.1.14 638223240 chitinase family 18 (EC:3.2.1.14) 638208607
ec_3_2_1_14 EC:3.2.1.14 638240603 chitinase family 18 (EC:3.2.1.14) 638208605
ec_3_2_1_14 EC:3.2.1.14 638262959 chitinase family 18 (EC:3.2.1.14) 638208601
ec_3_2_1_14 EC:3.2.1.14 638265193 chitinase 638208601
ec_3_2_1_14 EC:3.2.1.14 638265450 class V chitinase Chi100 638208601
ec_3_2_1_14 EC:3.2.1.14 638282069 chitinase family 18 (EC:3.2.1.14) 638276533
ec_3_2_1_14 EC:3.2.1.14 638288817 chitinase family 18 (EC:3.2.1.14) 638276032
ec_3_2_1_14 EC:3.2.1.14 638292698 chitinase family 18 (EC:3.2.1.14) 638276391
ec_3_2_1_14 EC:3.2.1.14 638296053 chitinase family 18 (EC:3.2.1.14) 638276690
ec_3_2_1_14 EC:3.2.1.14 638315678 chitinase familv 18 (EC:3.2.1.14) 638275801

```

You can create a gene list from the above file using Excel. Load the text file to Excel and delete all the columns except the gene OID (i.e., 3rd column). Then insert a row at the very top, and type in “gene” in the 1st column of this new first row, and type “ec_3_2_1_14” in the 2nd column. Output the Excel data to a text file, and you will have:

```
list - Notepad
File Edit Format View Help
gene\t ec_3_2_1_14
637099575
637163698
637165689
637205987
637268823
637269652
637400993
637455106
637503596
637790575
637790800
637831538
637842071
637907992
638001572
638024292
638032487
638119440
638147373
638223240
638240603
638262959
638265193
638265193
Ln 1, Col 17 100% Windows (CRLF) UTF-8
```

Note that “gene” and “ec_3_2_1_14” are separated by a tab.

Now go to **IMG UI Workspace** -> **Gene Sets** and click the “**Import & Export**” tab. In the **Import** section, choose the gene OID file you just created, and click the **Import Gene Sets** button, and you will have a new gene set **ec_3_2_1_14** in your workspace.

Step 3: Download all gene information

You are finally ready to download gene information.

Click to select gene set **ec_3_2_1_14** from Workspace Gene Sets:

Gene Workspace Tips ?

Gene Cart

You have [20000 Gene\(s\)](#) in your cart.

Gene Sets

Group sharing is not displayed. Please go to [Preferences](#) to change the sharing display options.

hint: Select one or more gene sets to perform gene set analysis. Click on the gene set count to view and analyze genes in a particular gene set.

Gene Sets Import & Export Genomes & Scaffolds Function Profile Set Operation

1 of 2 rows selected

Filter column: File Name Filter text Apply ?

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	File Name	Number of Genes (click the link to each individual set)	File Size
<input type="checkbox"/>	4_hydrob	43	462 B
<input checked="" type="checkbox"/>	ec_3_2_1_14	47082	503.690 KB

Export Page 1 of 1 << first < prev 1 next > last >> All

1 of 2 rows selected

Add Selected to Public

Add Selected to Gene Cart

Select All

Clear All

Remove Selected

Then go to the **Import & Export** tab again, and scroll down to find the **Data Export** section:

Gene Sets **Import & Export** Genomes & Scaffolds Function Profile Set Operation

Import

You may import gene sets from a file created by using the export feature below. A file can also be successfully imported if it follows a [specific format](#).
(Special characters in set name will be removed and spaces converted to _)

File to upload:
 No file chosen

Export

You may select one or more gene sets from above to export. The exported file may be imported later into your workspace.
NOTE: Exported gene sets contain IDs only. To export the contents of a gene set, please go to the gene set page.

Data Export

hint: Export large number of genes will be very slow.
 You will be notified for the result via email if exporting over 100 genes.

You may export data from the selected gene set(s).

My Email
 (Results will be mailed to you if selection is over 100 genes.)

Make sure that your email is shown up (or filled in) in the **My Email** field; otherwise you won't receive an email notification.

- **Fasta Nucleic Acid File:** download gene nucleic acid sequence
- **Fasta Amino Acid File:** download gene protein sequence
- **Gene Data In Excel:** download gene information in Excel

When the job is done, you will receive an email notification. Follow the instruction in the email to download your result file. The following is an example gene protein sequence file:

>637003882 chitinase (EC:3.2.1.14)
MKNLIFTRKSMIGMLVCSALPALAMEAWNQQGGNKYQVIFDGKIYENAW
WVSSSTNCPGKAKANDATNPWRLKRTATAAEISQFGNTLSCEKSGSSSSSN
SNTPASNTFPANGGSATPAQGTVPSNSSVAVWVKQGGQWYVVFNGAVYK
NAWVWASSNCPGDAKSNDAENPWRYVRAATATEISETSNPQSCTSA PQPS
PDVKPAPDVK PAVDVQ PAVADKSNNDNYAVVAWKGQEGSSTWYVIYNGGIY
KNAWVWGAANCPGDAKENDASNPWRYVRAATATEISQYGNPFGSCSVKPDN
NGGAVTPVDPTPETPVTPTPDNSEPSTPADSVNDYSLQAWSGQEGSEIYH
VIFNGNVYKNAWVWVSGKDCPRGTS AENSNNPWRLERTATAAELSQYGNPT
TCEIDNGGVIIVADGFQASKAYSADSIVDYNDAHYKTSVDQDAWGFVPGGD
NPWKYEPAKAWSASTVYVKGDRVVVDGQAYEALFWTQSDNPALVANQNA
TGSNSRFPWKPLGKAQSYSNEELNNAQFNPETLYASDTLIRFNGVNYISQ
SKVQKVS PDSNPWRV FVDWTGTKERVGT PKKAWPKHVYAPYVDFLNTI
PDLAALAKNHNHVNHTLAFVVS KDANTCLPTWGTAYGMQNYAQYSKIKAL
REAGGDVMSLIGGANNAPLAASCKNVDDLQHYDYDIVDNLNLKVLDFDIE
GTWVADQASIERRNLA VKKVQDKWKSEKGDIAIWYTLPIPTGLTPEGMN
VLSDAKAKGVELAGVNVMTMDYGN AICQS ANTEGQNIHGKCATSAIANLH
SQLKGLHPNKSDAEI DAMMGTTPMVGVNDVQGEVFYLS DARLVMQDAQKR
NLGMVGIWISIARDLPGGTNLSPEFHGLTKEQAPKYAFSEIFAPFTKQ

>637047549 chitinase, putative
MFNKISLEIMKRSALPLMPTVLALAVGMAMPVQAAINS DASVVGTESQW
WNTYKVTLTNNGNQPV ELRDASIAFDTNLSLSTPSWSAQQISYPSMSFSS
NAQGSVFSNRLTLSFDQGSWVK TQLLPGASIDLTLGVSGVLELSLQSTI
ALETDGEVEPEGEPEISLELASPVQGAEFIEGQTVAIVANVTATNTTVKTV
TFLVDGEQIALLEQAPFQASWTAAGEGAHSIKAIVEDASGLLKEQAVRIT
VKAEEIDPPVEPEVPVAPVIELTNP RNQSVFLGKVTTLAANATDENNDL
TAVEFLVNGESIGRVTQAPYQLAWTPVALGQYTI EAIAYDAAGHOTQTPM
VTVNAKEMGTGNLSCDIKQIYREDGTECMGDDHPRRIIGYYSWRGTGKNG
LPAYLAGDLPWEKLTHINYAFASINKSDFSMQVDDSATKMTWENVPGAEM
DPSLPYQGHFNLLSKFKKQYPDVKT LISVGGWAETGGFYPM TDLASCSV
NMEGIKAFNKSAVD FIRQYDFDGV DIDIYEY PSSMKDSGNPVDFEQSNKCR
GQLWDNYMVMTELRKALDKAGEEDGRRYMLTIAS PSSAYLLRGMQDFAM
QDVL DYVNIMS YDLHGTWNEFVGFQAALFDDGKDAELAKWGVYTTAEYQG
IGYLNQAWTHHFFRGA FKPQSINMGI PYYTRGWQGVSGGDKGLWGRAVEP
NQSSCEGT TVCGWGAEGTDNIWHDVDANGNEIKAGVVP MWHAMNLMHAE
KLGIDGMPSYGPWGMDFNNPKHLIEGKYERVWSQELQTAWLWNDTKKVF
LSIEDKDSLKPKLDYIVDNLGLGMMIEMAGDYSYDAVKREYVIGSDMTS