Exercise: Ensembl Fungi BioMart
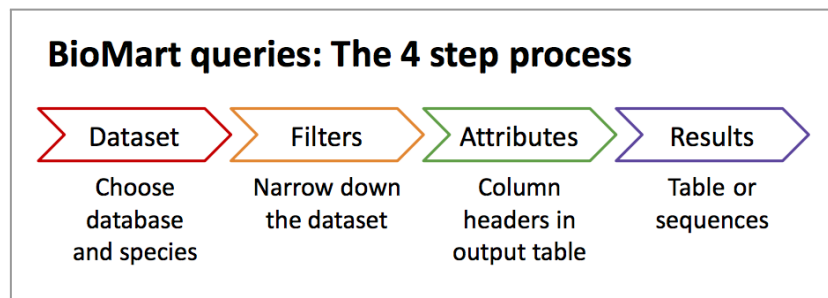
Follow these instructions to guide you through BioMart to answer the following query:

(a) How many genes within the 14:1128520-1142558 region are found in *Fusarium solani* that do not have an orthologue in *Fusarium verticillioides*?
(b) Export the gene name, locations and GO terms associated with these genes
(c) Export their cDNA sequences

Click on BioMart in the top header of a fungi.ensembl.org page to go to:
https://fungi.ensembl.org/biomart/martview/

NOTE: These answers were determined using BioMart Ensembl Fungi 56

Step 1a: Choose Ensembl Fungi Genes 56 as the database

Step 1b: Choose *Fusarium solani* genes (v2.0) as the dataset

Step 2: Choose appropriate filters

We want to narrow down the dataset of all *F. solani* genes to a subset of genes matching our filters. We are interested in *F. solani* genes that **do not** have an orthologue with *F. verticillioides*. We need to filter the dataset to find these genes.
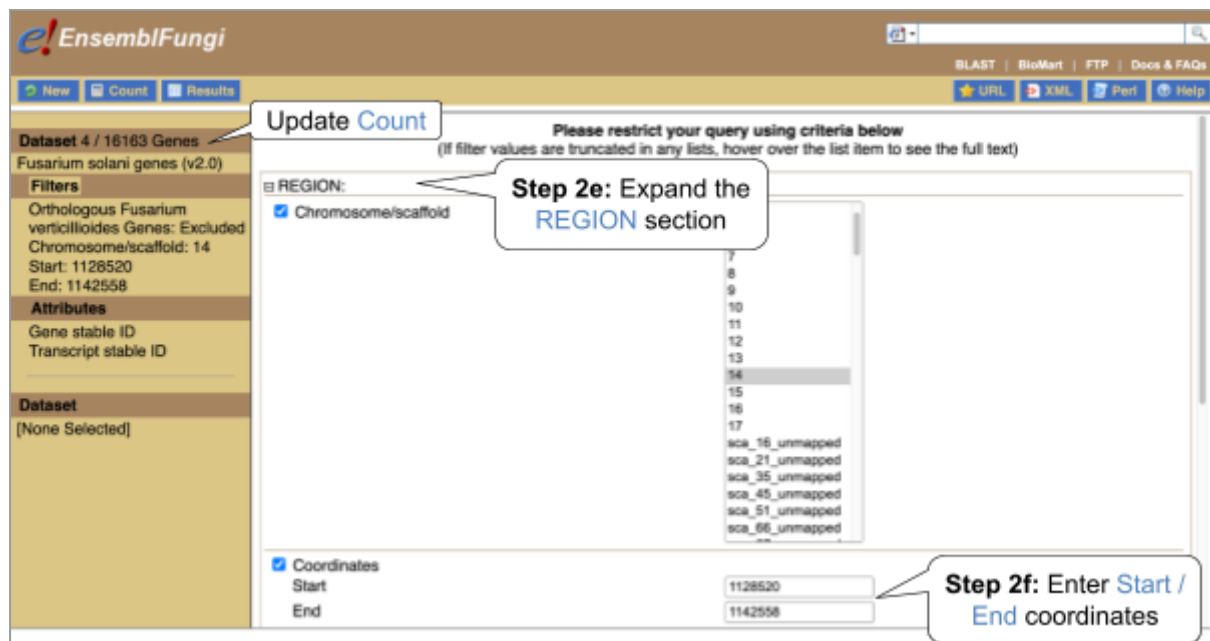
Using the count function we can see that there are 4 *F. solani* genes (out of a total of 16,163) in the 14:1128520-1142558 region that do not have an orthologue in *F. verticillioides*.

Step 3: Select Attributes

Attributes (our desired output) are defined by what we would like to learn about the data. We want to find out more information about these genes, including:

1. Gene name
2. Locations
3. Associated GO terms
4. cDNA sequences

There are four main attribute types: Features, Structures, Homologues and Sequences. BioMart allows querying only one type at a time. We can answer points 1-3 in a single query as they can all be found under Features, but we will need to build a second query to answer point 4 (Sequence type).

Make sure that Features is selected at the top of the page.
Expand the **GENE** section, and **select** Chromosome/scaffold name, Gene start and Gene end, and Gene name.

Expand the **EXTERNAL** section. This section contains lots of identifiers from databases outside of Ensembl. Select GO term accession and GO term name.

Step 4: Get results!

You can download the data if you'd like. The output table shows only 10 first rows by default.

You can click on the location links and explore the synteny between the two species.

What about the last point? 'Export their cDNA sequences?'
In the Attributes section there are some 'radio buttons'. If you'd like to export Sequence data, you need to build a separate query.

Step 3.2: Let's go back to step 3: Selecting attributes

From the results page, click back to Attributes in the left-hand navigation panel – there's no need to start from scratch.

Also expand the **HEADER INFORMATION** section and **select** Gene name.

Step 4.2: View results for the sequences

What did you learn about these genes in this exercise?
Could you learn these things from the Ensembl browser? Would it take longer?

For more details on BioMart, have a look at this publication:

Kinsella RJ, Kähäri A, Haider S, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database : the Journal of Biological Databases and Curation. 2011;2011:bar030. DOI: 10.1093/database/bar030. PMID: 21785142; PMCID: PMC3170168.

Additional BioMart Exercise 1 – Export orthologues

Use Ensembl Fungi BioMart to retrieve all *Zymoseptoria tritici* genes associated with the GO term detoxification located on chromosome 1. Export the gene IDs, names, homology type and confidence of their orthologues in *Blumeria graminis*, *Botrytis cinerea*, *Cryptococcus neoformans* and *Saccharomyces cerevisiae*.

   (a) Do all of these *Z. tritici* genes have an orthologue in the other species? Which of these species are pathogenic? Do you see a correlation?

   (b) Can you find an orthologue in *Cryptococcus neoformans* with high orthology confidence? What is the Gene ID? We will explore more about this orthologue in the exercise section for the Evolutionary Analysis module.

Dataset 18 / 11091 Genes
Zymoseptoria tritici genes (MG2)

Filters

Chromosome/scaffold: 1
GO Term Name [e.g. regulation of biological process]: detoxification

Attributes

Gene stable ID
Transcript stable ID
Gene name
Blumeria graminis gene stable ID
Blumeria graminis gene name
Blumeria graminis homology type
Blumeria graminis orthology confidence [0 low, 1 high]
Botrytis cinerea B05.10 gene stable ID
Botrytis cinerea B05.10 gene name
Botrytis cinerea B05.10 homology type

Export all results to [File ▾] [TSV ▾] ☐ Unique results only [● Go]

Email notification to [_____]

View [10 ▾] rows as [HTML ▾] ☐ Unique results only

Gene stable ID	Transcript stable ID	Gene name	Blumeria graminis gene stable ID	Blumeria graminis gene name	Blumeria graminis homology type	Blumeria graminis orthology confidence [0 low, 1 high]	Botrytis cinerea B05.10 gene stable ID	Botrytis cinerea B05.10 gene name	Botrytis cinerea B05.10 homology type	Botrytis cinerea B05.10 orthology confidence [0 low, 1 high]
Mycgr3G90067	Mycgr3T90067									
Mycgr3G96385	Mycgr3T96385						Bcin16g00120	Bcprx3	ortholog_one2one	0
Mycgr3G33131	Mycgr3T33131									
Mycgr3G107202	Mycgr3T107202		BLGH_06273		ortholog_one2one	0	Bcin10g02340		ortholog_one2one	0
Mycgr3G54449	Mycgr3T54449		BLGH_05551		ortholog_one2one	0	Bcin14g01030	Bcsfa1	ortholog_one2one	0
Mycgr3G32802	Mycgr3T32802						Bcin05g00590	Bcprd4	ortholog_one2many	0
Mycgr3G32802	Mycgr3T32802						Bcin05g01450	Bcprd2	ortholog_one2many	0
Mycgr3G24719	Mycgr3T24719									

Additional BioMart Exercise 2 – Finding genes by protein domain

Generate a list of all *Magnaporthe oryzae* genes on chromosome 4 that are annotated to contain Transmembrane domains/helices. Include the Ensembl Gene ID and description.

Additional BioMart Exercise 3 – Convert IDs

For a list of *Schizosaccharomyces pombe* UniProt (UniProtKB/Swiss-Prot) IDs, export the Gene name and description, as well as the PomBase IDs. Do these 36 protein IDs correspond to 36 genes?

Input list of IDs:

Q92338	Q9US55	P78847	O74964
O13728	O14075	O94418	O14026
P49776	O94574	O94526	O74630
O74769	O94380	Q9UTG2	O14356
Q09170	P87172	O14326	O13339
Q9USK4	Q9USP5	Q9URZ3	P31411
O14040	Q9P7Y8	P42657	O13742
Q9Y804	Q9Y7Z8	P08647	O60159
O94552	Q10331	O74335	O94287

Dataset 36 / 7268 Genes

Schizosaccharomyces pombe genes (ASM294v2)

Filters

UniProtKB/Swiss-Prot ID(s) [e.g. A0ZWU1]: [ID-list specified]

Attributes

Gene stable ID
Gene name
Gene description
PomBase ID

Dataset

[None Selected]

Export all results to | File ⬍ | HTML ⬍ | ⎵ Unique results only ⬤ Go

Email notification to

View | 50 ⬍ | rows as | HTML ⬍ | ⎵ Unique results only

Gene stable ID	Gene name	Gene description	PomBase ID
SPBC29A3.14c	trt1	telomerase reverse transcriptase 1 protein Trt1 [Source:PomBase;Acc:SPBC29A3.14c]	SPBC29A3.14c.1
SPAC15A10.08	ain1	alpha-actinin [Source:PomBase;Acc:SPAC15A10.08]	SPAC15A10.08.1
SPAC16E8.07c	vph1	V-type ATPase V0 subunit a (predicted) [Source:PomBase;Acc:SPAC16E8.07c]	SPAC16E8.07c.1
SPAC29B12.02c	set2	histone lysine methyltransferase Set2 [Source:PomBase;Acc:SPAC29B12.02c]	SPAC29B12.02c.1
SPAC2C4.07c	dis32	3'-5'-exoribonuclease activity Dis3L2 [Source:PomBase;Acc:SPAC2C4.07c]	SPAC2C4.07c.1
SPACUNK4.10		glyoxylate reductase (predicted) [Source:PomBase;Acc:SPACUNK4.10]	SPACUNK4.10.1
SPBC16E9.11c	pub3	HECT-type ubiquitin-protein ligase E3 Pub3 (predicted) [Source:PomBase;Acc:SPBC16E9.11c]	SPBC16E9.11c.1
SPBC30D10.10c	tor1	phosphatidylinositol kinase Tor1 [Source:PomBase;Acc:SPBC30D10.10c]	SPBC30D10.10c.1
SPBC19C7.11		ClC chloride channel (predicted) [Source:PomBase;Acc:SPBC19C7.11]	SPBC19C7.11.1
SPBC17F3.01c	rga5	Rho-type GTPase activating protein Rga5 [Source:PomBase;Acc:SPBC17F3.01c]	SPBC17F3.01c.1
SPCC23B6.03c	tel1	ATM checkpoint kinase [Source:PomBase;Acc:SPCC23B6.03c]	SPCC23B6.03c.1
SPBC24C6.08c	bhd1	folliculin/Birt-Hogg-Dube syndrome ortholog Bhd1 [Source:PomBase;Acc:SPBC24C6.08c]	SPBC24C6.08c.1
SPBC4B4.03	rsc1	RSC complex subunit Rsc1 [Source:PomBase;Acc:SPBC4B4.03]	SPBC4B4.03.1
SPBC887.15		ClC chloride channel (predicted) [Source:PomBase;Acc:SPBC887.02]	SPBC887.02.1
SPBC1604.15	gpi16	pig-T, Gpi16 (predicted) [Source:PomBase;Acc:SPBC1604.15]	SPBC1604.15.1
SPCC1620.11	nup97	nucleoporin Nic96 homolog [Source:PomBase;Acc:SPCC1620.11]	SPCC1620.11.1
SPBC609.02	ptn1	phosphatidylinositol-3,4,5-trisphosphate3-phosphatase Ptn1 [Source:PomBase;Acc:SPBC609.02]	SPBC609.02.1
SPCC18.18c	fum1	fumarate hydratase (predicted) [Source:PomBase;Acc:SPCC18.18c]	SPCC18.18c.1
SPBC1773.17c		glyoxylate reductase (predicted) [Source:PomBase;Acc:SPBC1773.17c]	SPBC1773.17c.1
SPAC17H9.09c	ras1	GTPase Ras1 [Source:PomBase;Acc:SPAC17H9.09c]	SPAC17H9.09c.1
SPAC637.05c	vma2	V-type ATPase V1 subunit B [Source:PomBase;Acc:SPAC637.05c]	SPAC637.05c.1
SPAC17A2.13c	rad25	14-3-3 protein Rad25 [Source:PomBase;Acc:SPAC17A2.13c]	SPAC17A2.13c.1
SPCC4G3.02	aph1	bis(5'-nucleosidyl)-tetraphosphatase [Source:PomBase;Acc:SPCC4G3.02]	SPCC4G3.02.1
SPCC290.03c	nup186	nucleoporin Nup186 [Source:PomBase;Acc:SPCC290.03c]	SPCC290.03c.1
SPBC3D6.07	gpi3	pig-A, phosphatidylinositol N-acetylglucosaminyltransferase subunit Gpi3 (predicted) [Source:PomBase;Acc:SPBC3D6.07]	SPBC3D6.07.1
SPCC18B5.11c	cds1	replication checkpoint kinase Cds1 [Source:PomBase;Acc:SPCC18B5.11c]	SPCC18B5.11c.1
SPBC428.01c	nup107	nucleoporin Nup107 [Source:PomBase;Acc:SPBC428.01c]	SPBC428.01c.1
SPBC2D10.18	abc1	ABC1 kinase family ubiquinone biosynthesis protein Abc1/Coq8 [Source:PomBase;Acc:SPBC2D10.18]	SPBC2D10.18.1
SPAPYUG7.03c	mid2	medial ring protein Mid2 [Source:PomBase;Acc:SPAPYUG7.03c]	SPAPYUG7.03c.1
SPAC869.10c	put4	proline specific plasma membrane permease Put4 (predicted) [Source:PomBase;Acc:SPAC869.10c]	SPAC869.10c.1
SPAC1002.03c	gls2	glucosidase II alpha subunit Gls2 [Source:PomBase;Acc:SPAC1002.03c]	SPAC1002.03c.1
SPCC4B3.14	cwf20	complexed with Cdc5 protein Cwf20 [Source:PomBase;Acc:SPCC4B3.14]	SPCC4B3.14.1
SPCC11E10.02c	gpi8	pig-K [Source:PomBase;Acc:SPCC11E10.02c]	SPCC11E10.02c.1
SPAC1805.15c	pub2	HECT-type ubiquitin-protein ligase E3 Pub2 [Source:PomBase;Acc:SPAC1805.15c]	SPAC1805.15c.1
SPBC146.13c	myo1	myosin type I [Source:PomBase;Acc:SPBC146.13c]	SPBC146.13c.1
SPBC146.06c	fan1	Fanconi-associated nuclease Fan1 [Source:PomBase;Acc:SPBC146.06c]	SPBC146.06c.1