Fat&MuscleDB: A database to understand tissue growth processes contributing to body or muscle composition

Jérémy Tournayre, Isabelle Cassar-Malek, Matthieu Matthieu.Reichstadt@inrae.Fr Reichstadt, Brigitte B. Picard, Nicolas Kaspric, Muriel Bonnet

Fat&MuscleDB: A database to understand tissue growth processes contributing to body or muscle composition

Jérémy TOURNAYRE[1,2], Isabelle CASSAR-MALEK[1,2], Matthieu REICHSTADT[1,2], Brigitte PICARD[1,2], Nicolas KASPRIC[1,2] and Muriel BONNET[1,2]

[1] INRA, UMR1213 Herbivores, Site de Theix, 63122, Saint-Genès-Champanelle, France

[2] Clermont Université, VetAgro Sup, UMR1213 Herbivores, BP 10448, 63000, Clermont-Ferrand, France

Corresponding authors: jeremy.tournayre@clermont.inra.fr and muriel.bonnet@clermont.inra.fr

Abstract *To minimise unnecessary redundancy in research efforts by a better use of available data, we present a web-based database and data-mining platform named Fat&MuscleDB. Genomics on muscle and adipose tissue growth has generated huge amount of data which are available in journals and in databases. Unfortunately these data are scattered on the Internet in a heterogeneous format. Thus, it is difficult to exploit them efficiently. We hypothesise that these data can allow identifying genes or proteins involved in adipose and muscle tissues development contributing to body or muscle composition, two key criteria of carcass and meat quality. Currently, Fat&MuscleDB contains genomic expression data and differential abundance data from about 100 publications and 75 GEO datasets. These data can be queried, visualised, and downloaded in different ways: the data visualisation of each reference, the search of transcripts or proteins in references, and the data aggregation based on criteria of adipose and muscle growth. The aggregation function of Fat&MuscleDB is illustrated through two questions: "What are the proteins secreted by muscles?" and "What are the transcripts and proteins involved in the growth of muscle tissue from genetic origins of bovine?".*

Keywords database, adipose tissues, skeletal muscle tissues, genomic expression

Fat&MuscleDB: Une base de données pour comprendre les processus de croissance des tissus contribuant à la composition corporelle ou musculaire

Résumé en Français : *Afin de minimiser les efforts de recherches inutiles par une meilleure utilisation des données disponibles, nous vous présentons une base de données en ligne pourvue d'un outil d'agrégation de données appelée Fat&MuscleDB. La génomique utilisée pour comprendre les mécanismes de la croissance des tissus musculaires et adipeux a généré une énorme quantité de données disponibles dans des revues et des bases de données. Malheureusement, ces données sont dispersées sur l'Internet dans un format hétérogène. Ainsi, il est difficile de les exploiter efficacement. Nous émettons l'hypothèse que ces données peuvent permettre d'identifier des gènes ou des protéines impliqués dans le développement des tissus adipeux et musculaires contribuant à la composition corporelle ou musculaire, deux critères clés de la qualité des carcasses et des viandes chez les ruminants. Actuellement, Fat&MuscleDB contient des données d'expression génique et des données d'abondance différentielle extraites d'environ 100 publications et 75 ensembles de données de GEO. Pour interroger, visualiser et télécharger ces données les outils proposés sont : la visualisation des données par référence, la recherche de transcrits ou de protéines dans les références et l'agrégation de données basée sur des critères de croissances adipeuse et musculaire. La fonction d'agrégation de Fat&MuscleDB est illustrée à travers deux questions: "Quelles sont les protéines sécrétées par les muscles?" et "Quels sont les transcrits et les protéines impliqués dans la croissance du tissu musculaire en fonction des origines génétiques des bovins?".*

Mots-clés *base de données, tissus adipeux, tissus musculaires squelettiques, expression génomique*

1. Introduction

Increasing the amount of meat produced while preserving a high quality is an economic challenge for the beef industry. The characterisation of meat quality depends on several criteria: tenderness, juiciness, texture, and flavour. These criteria are related to the amount of muscle tissue relative to the amount of adipose tissue: the lean-to-fat ratio. Thus, many studies have been done on the changes of this ratio according to different

animal breeds and animal husbandry practices. Genomic studies have identified genes and proteins which control tissue physiology and growth. These genomic data are published in many journals available in public databases such as PubMed [1] and Web of Science (www.webofknowledge.com) whose the transcriptomic data can be filed into the Gene Expression Omnibus (GEO) database [2]. Some people have developed databases that aim to integrate these data in other contexts of research. For example, databases like Oncomine [3] and EURRECA [4] are dedicated to identification of cancer biomarkers and micronutrient status respectively. Concerning our topics, several published transcriptomic data from the muscle were processed in the MADMUSCLE [5] database, but it is not maintained and not updated anymore. We can also use the StemBase [6] database which lists transcriptomic data from the muscle and adipose stem cells from 62 experiments, without any access to lists of genes absent, present or differentially expressed between two samples. Thus, there is no database that aims at integrating transcriptomic and proteomic data for muscle and adipose tissues growth. The integration and the aggregation of these data could help answering to other questions than those raised by their authors, e.g. the emerging issue: what are the developmental and functional links between muscle and adipose growth? This question is suggested by the successive waves of muscle and adipose tissues growth as well as the wide plasticity of body composition depending on age or genotype in cattle [7]. Currently, little functional links are reported in monogastrics thanks to *in vitro* studies [8].

Our objective is to retrieve all the data which could help us to understand the mechanisms underlying fat and muscle hypertrophy and hyperplasia. In order to achieve this, we created a web-based database and data-mining platform called Fat&MuscleDB. The data come from *in vivo* and *in vitro* experiments in ruminants and model species such as rodents, humans, and cell lines. They were collected at various ages since muscular hypertrophy and fat hypertrophy occur during foetal life and in postnatal life. We defined criteria relative to the experiments and the stages of adipose tissue and muscle growth in order to record and classify data. Our final aim is their aggregation to highlight transcripts or proteins which may be biomarkers of the lean-to-fat ratio. To construct Fat&MuscleDB, a data collection was created and was semi-automatically processed to extract and import the data in the database. We developed an interface which provides the view of all data, the search of transcripts or proteins in the database, and the aggregation based on classification criteria.

2. Methods

a. Data collection

For the collect of data on our topics, we listed keywords related to cellular and tissue traits linked to muscle and adipose tissues growth, species and cell lines, methods, and keywords for irrelevant references exclusion. All combinations of these keywords were submitted to NCBI on GEO and Pubmed, and also on Web of Science. The querying of Pubmed and GEO databases on the NCBI was made with Perl and the LWP::Simple library (http://search.cpan.org/~ether/libwww-perl-6.13/lib/LWP/Simple.pm), whereas querying Web of Science was done with Selenium Remote-Control (http://www.seleniumhq.org/) controlled by a Perl script which allowed us to use the Firefox browser in order to interpret the JavaScript language. To avoid overload of the queried servers, we used the NCBI web service and put a waiting time between the display of each page of Web of Science without parallelisation. We created our own website to view and manipulate this data collection.

b. Data processing

We extracted data from publications and supplementary data to retrieve all genes or proteins names together with the fold change. For the extraction of these columns from Portable Document Format (PDF) two tools was used: Tabula (www.tabula.technology) and pdftotext (www.glyphandcog.com). By this way, the genomic expressions were extracted and classified as "Absence" or "Presence". In addition, the differentially abundant genes or proteins between two samples were extracted and classified as "Increased", "Decreased" or "Stable". For data related to the secretome assayed with cell culture the authors may have used tools such as SignalP [9] to eliminate results from contamination due to cell death and to identify accurately proteins that are secreted. If this was not done, SignalP was used from the website ProteINSIDE (www.proteinside.org, see abstract from N. Kaspric) to remove falsely predicted proteins which were the result of a contamination.

On the other hand, intensities of transcripts were extracted from GEO microarray data. Here we wanted to identify differentially expressed genes between two samples. They can be together on the same chip (two channels) or on separate chips (one channel or two channels with a reference on the second channel). Each microarray dataset was verified manually to take into account if there was a dye swap or a dye switch, if the data were paired, if the data were normalised and if the data were log-transformed. An R script which took into account these indications was made. This script downloaded GEO data as a data.frame thanks to the GeoQuery library [10]. Then, for each type of chip used, according to our indications, the script normalised values, converted them into log base 2 and reversed the signs of the values in the case of a dye swap or dye switch. Then, the values of replicated probes for each chip were averaged as well as the microarray technical replicates. The column indicating the probes name was selected automatically but could be handpicked manually if the selected column was not correct. Indeed, this column does not have the same name and was not at the same location in each microarray dataset. However, we were able to set some rules to address automatically the probes column: if a column was named "ReporterID", if a column had no missing value and had no more than 15% of identical identifiers (which may correspond to a "block", "row", or "column" of a chip) and the average size of string was less than 30 (above this number it may be sequences or descriptions of transcripts). Among these columns the script chose the column whose identifiers ending with "_at" (this is the end of an identifier of Affymetrix chip), or the column whose name was "INT_ID" and strings began with "C" (this is a clone identifier of GenBank), or the column which had the most unique identifiers among all. To finish and to identify differentially expressed transcripts, a variance mixture was performed thanks to the R library Anapuce (http://cran.rproject.org/web/packages/anapuce/) which is dedicated to microarray data analysis. The "DiffAnalysis" function was chosen if the data were paired. It took into account only samples on a same chip; so if samples were biologically paired and were on different microarrays the R script calculated each value of the first sample minus the second sample. The "DiffAnalysis.unpaired" function was selected for the unpaired data. These functions allowed obtaining over-expressed, under-expressed, and stable genes lists. A Benjamini and Hochberg *p*-value threshold of 0.15 was chosen because the stringency of this test strongly reduced the number of regulated transcripts identified. We slightly changed the Anapuce functions to solve rare bugs and to only perform the variance mixture. For example, an error was caused if the value of Bayesian Information Criterion was greater than 10^7 in the first iteration of variance mixture. Also, the "lowess" function was stopped if there were missing values in the vector. We also treated genomic expressions data within a sample. For that, genes were classified in "Absence" or "Presence" if for each chip the probe was recorded as "A" or as "P" respectively.

We chosen to distinguish the data analysed by our methods (e.g. by Anapuce or a SignalP analysis) of the data extracted as they were published. The firsts were annotated by "Analysed by Fat&MuscleDB" while the others were annotated by "Not analysed by Fat&MuscleDB". To aggregate these data it was necessary to have a unique identifier for each imported data. Thus, all names or identifiers were chosen to be converted into UniProt accessions from UniProt [11] which is a protein database regularly updated and curated by experts. For this, we chose the names to be converted according to the species studied using the UniProt or NCBI search engine depending on their nomenclature. We retrieved the "Entry" (e.g. "Q15848"), the "Gene names" (e.g. "ADIPOQ"), the "Entry name" (e.g. "ADIPO_HUMAN") and the information that the accession was reviewed by UniProt or not for each UniProt accession. This was done automatically through a Perl script using the tabulated format of the UniProt research results for only get the desired information (e.g.: "http://www.uniprot.org/uniprot/?query=adipoq+AND+organism%3Ahuman&sort=score&limit=10&format=tab&columns=id,entry%20name,reviewed,genes"). Finally, the aggregations were performed on the "Gene names" because these accessions were not linked to species contrary to the "Entry name" and the "Entry". To classify all extracted data we defined criteria related to muscle or adipose tissues growth from both *in vivo* and *in vitro* experiments and criteria related to species, breeds, genders, ages, tissues, and cell lines. These criteria were selected for each extracted data based on the information described in the experimental protocols indicated by the authors. To finish, these data were imported in a MySQL database.

c. Web interface

The web interface was programmed in PHP, HTML, and JavaScript, which means that it works on the most used browsers (kept up to date) (Chrome, Firefox, Internet Explorer...). Tables on the web interface were created using Google Charts (developers.google.com/chart) that enables to create easily interactive tables with search field. The website offers three types of searches: viewing all data, searching for a list of genes or

proteins in the database, and aggregation of the data based on the criteria of hypertrophy or atrophy in muscle or adipose tissues as defined in the classification. Each search results can be downloaded in Excel format (.xlsx). The aggregation was processed by a Perl script on the user request. This produced a table that merged results from all experiments sharing a same criterion. Each accession was counted once by reference in order to identify the redundancy of identification of a gene or a protein. Each aggregation result produced a code for easy retrieval in the web site. The search for genes or proteins lists can be done with these UniProt accessions: "Entry", "Entry name" or "Gene names". We used a denormalised table where each UniProt accessions was connected to their reference to offer a quick search among all the data. Data imported from references and aggregation results were precalculated for a quick display of data on a web page and also a quick download of Excel files.

3. Results and discussion

a. Data collection

To create the database collection, references were searched on Web of Science, Pubmed, and GEO databases. This research was done automatically using a combination of three lists of keywords: 86 keywords describing cellular and tissular traits linked to muscle and adipose tissues growth (including adipose tissue, muscle, marbling, double-muscled, carcasses, meat qualities...), 26 on species and cell lines (bovine, foetus, 3T3-L1, C2C12...), and 12 on methods (transcriptome, proteome...). Furthermore, a list of negative keywords (diseases, carcinoma...) has been added on each of these combinations in order to discard data from abnormal conditions of tissues growth. This generated about 26,000 queries on each database which allowed us to retrieve about 17,000 publications and 2,500 GEO. We currently view these references through our private website to facilitate the selection of the most relevant studies about muscle and adipose tissues growth. Presently, we validate 345 publications and 162 GEO datasets to extract. We reject about 4,500 publications and about 1,200 GEO datasets for multiple reasons: they may not contain table, they can deal with a too specific tissue growth (for example, a knockdown of a gene contributed to adipose or muscle tissues growth). Also, the NCBI and Web of Science search engines retrieved publications related to diseases despite the list of negative keywords. So, we automatically reject references which contain, in their title, one of the negative keywords without positive keywords like "adipo*". More than 2,000 references have been rejected by this way.

For the moment, we extract and import into the database about 100 publications and about 75 GEO datasets. The time of analysis is substantial because we want to classify data accurately which require a good understanding of the material and methods. During the data classification it happened that finally we rejected the reference after careful consideration on the methods or the data. For example, sometimes there were no biological replicates in a GEO dataset which prevented us from achieving statistical analysis of variance mixture. Mistakes were often found in the literature: e.g. the age of one animal is different between summary and methods, a gene in a table can be specified as over-expressed and also described under-expressed later in the same table for the same sample... Also, it is difficult to accurately extract the data automatically from publications because each researcher put his own nomenclature: for example, for column header title indicating gene names we can have "Accession", "Gene", "ID"... or even no title. In a similar way as publications there is heterogeneity in the description of GEO transcriptomic data despite of the NCBI rules. For example, the intensity values may be normalised or not, in a log base 2 or in a log base 10 or even not in a logarithmic form, the dye swaps are indicated in different ways, sometimes there are uninformative terms e.g. a date can be specified only by "time 1", "time 2"... We identify data from RNA-seq which are represented by about 200 GEO datasets in our data collection. In a second step, it might be interesting to take them into account with a statistical method, such as given by the R libraries DESeq [12] or NOIseq [13]. In order to aggregate the data thanks to unique identifier, the names given by the authors are converted into "Entry", "Entry name", and "Gene Names" of UniProt through UniProt or NCBI search engines (depends on the used nomenclature). During this conversion we retrieve the "Gene names" with some troubles. UniProt uses space as delimiter field for the "Gene names" which causes poor retrieval because few "Gene names" can contain spaces. For example, the "Entry" Q91VB8 have these "Gene names": "Hba-a1", "haemaglobin alpha 1", "haemaglobin alpha 2", and "Hba-a2" (see: "http://www.uniprot.org/uniprot/?query=Q91VB8&sort=score&limit=10&format=tab&columns=id,entry%20name,reviewed,genes" (this search was verified the 03/17/2015)). Thus, "haemaglobin alpha 1" is retrieved as "haemaglobin", "alpha", and "1" in our database. Consequently, these rare genes names cannot be searched in

the database and they are displayed and aggregated incorrectly in the results files. We contacted the UniProt staff to ask them if they could solve this field separator problem. They answered us that indeed there is a field separator problem and that it could take a lot of time to be resolved because they have to find a suitable field separator. The data heterogeneity, the data errors and the lack of precision on the data context as we have shown were already raised [14]. Communities propose to improve the sharing of biological data by defining standards, such as the standard "Minimum Information About a Microarray Experiment" (MIAME) (www.fged.org) which is required to deposit microarray data in GEO.

b. Classification

Up to now, we have defined 83 criteria to classify experiments and stages of adipose and muscle tissues growth. These criteria are grouped as: "*In Vitro*", "*In Vivo*", and "Others" main criteria. In "*In Vitro*" we decided to distinguish the mouse cell lines (3T3-L1, C2C12, C3H10T/2) from the primary foetal cells (Embryonic Stem Cell) and adult cells (Adult Mesenchymal Stem Cell, Adult Mesenchymal Stem Cell Adipogenic, Adult Mesenchymal Stem Cell Myogenic, Adipogenic progenitors, Myogenic progenitors), and from the secretome. Each cell types were classified according to the different stages of proliferation or differentiation. In "*In Vivo*" we distinguished fat from muscle, hypertrophy from atrophy induced by genetic (breed, gender, and genetic mutation), diet, and also from the secretome described from biological fluids or predicted from tissues or cell cultures. Finally, we classified in "Others" the comparisons between anatomical site and cell types namely, brown, beige and white adipocytes, and myocytes.

First condition	Second condition	Molecule	State	Species	Breed	Biological sample	Cell	Chip(s)	Reference	Authors	File(s)
Muscular hypertrophy from genetic origin	Reference	Transcripts	Double muscled VS Control	Bovine	Charolais (260 dpc)	Semitendinosus muscle	[NA]	[NA]	Target genes of myostatin loss-of-function in muscles of late bovine fetuses. Link DOI Link Pubmed	Cassar-Malek I ; Passelaigue F ; Bernard C ; Léger J ; Hocquette JF	Increased Decreased

Decreased | Increased

Analysed by Fat&MuscleDB	Uniprot conversion	Entry Name	Gene Names	Reviewed	Gene Symbol	Gene name	Fold
No	Q3U9N9	MOT10_MOUSE	MCT10 - SLC16A10 - Tat1	Yes	2610103N14Rik	Slc16a10 Solute carrier family 16	2.76
No	Q29RH4	THOC3_BOVIN	THOC3	Yes	2410044K02Rik	Thoc3 THO complex 3	2.64
No	[NA]	[NA]	[NA]	[NA]	LOC58504	Hypothetical protein from clones 23549 and 23762	2.52
No	E1BMB2	E1BMB2_BOVIN	Slc26a4	No	SLC26A4	solute carrier family 26, member 4	2.48
No	I7G9K3	I7G9K3_MACFA	[NA]	No	FLJ13855	hypothetical protein FLJ13855	2.36
No	Q9Y236	OSGI2_HUMAN	OSGIN2 - C8orf1	Yes	C8orf1	Chromosome 8 open reading frame 1	2.27
No	G3X6W9	G3X6W9_BOVIN	Mybph	No	MYBPH	myosin binding protein H	2.27
No	Q13495	MAMD1_HUMAN	MAMLD1 - CXorf6 - CG1	Yes	CXorf6	chromosome X open reading frame 6	2.19

Figure 1. Visualisation of transcriptomic data from the publication "Target genes of myostatin loss-of-function in muscles of late bovine foetuses" [15] classified as "Muscular hypertrophy from genetic origin". This is an extract of the list of transcripts whose expression increases in double-muscled Charolais compared to normal-muscled Charolais as reference. Each line corresponds to a transcript and brings various information (from left to right): if it was retrieved by us ("Yes") or according to authors' interpretation ("No"), its UniProt accessions: "Entry", "Entry name", "Gene Names", if it were reviewed by the UniProt consortium or not. In the other columns (e.g. "Gene Symbol", "Gene name", and "Fold") the data were extracted from the publications. The list of transcripts with a decreased expression in double-muscled Charolais can be downloaded by clicking the "Decreased" button on the upper left. Each column can be filtered and sorted.

c. View

It is possible to view and download all the data of Fat&MuscleDB. For example, the transcripts from the publication "Target genes of myostatin loss-of-function in muscles of late bovine fetuses" [15] are classified as over- and under-expressed in double-muscled Charolais with the criterion "Muscular hypertrophy from genetic origin" by comparison with normal-muscled Charolais. In addition, to the data we add a column named

"Analysed by Fat&MuscleDB" which indicates whether the data were retrieved as described in the publication or if we used a variance mixture from Anapuce or another of our methods. Also, the result of UniProt conversion is given in the columns "UniProt conversion", "Entry Name", "Gene Names", and "Reviewed" of the output table (Figure 1).

d. Search accession

In order to find publications and GEO datasets where a gene or a protein are identified, the search of accessions is possible through the following types of identifiers: "Gene name", "UniProt accession" or "Entry Name". The example of adiponectin with the "ADIPOQ" gene name search shows, as expected, that the protein is present among the proteins secreted by adipose tissues in rats and humans (Figure 2). Also, it is found increased in adipose hypertrophy induced by diet and by growth, and is decreased in intramuscular fat relative to omental adipose tissue.

You searched for

ADIPOQ

Presence

First condition	Second condition	Molecule	State	Species	Breed (age)	Biological sample	Cell (age)	Chip(s)	Reference	Authors
Proteins secreted by adipose tissues	Proteins secreted by adipose tissues	Proteins	[NA]	Rat	Sprague-Dawley Male (5 Weeks)	Gonadal adipose tissue	[NA]	[NA]	Secretome analysis of rat adipose tissues shows location-specific roles	Roca-Rivada, Arturo ; Alonso, Jana ; Al-Massadi, Omar ; Castelao, Cecilia ; Ramon Peinado, Jua Maria Seoane. Luisa
Proteins secreted by adipose tissues	Proteins secreted by adipose tissues	Proteins	Undergoing laparotomy to remove an intramural myoma	Human	(39 Years)	Omental	[NA]	[NA]	Comparison of isotope-labeled amino acid incorporation rates	Roelofsen H ; Dijkstra M ; Weening D ; de Vries MP ; Hoek A ; Vonk RJ

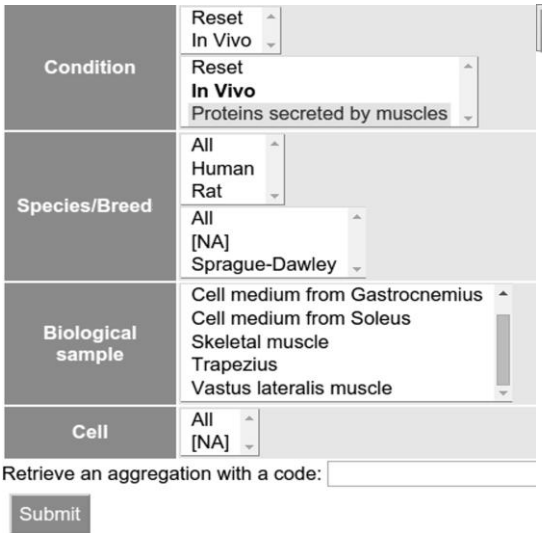
Figure 2. Visualisation of references which contain ADIPOQ in their dataset. This is an extract of the list of references retrieved from the search of ADIPOQ in Fat&MuscleDB. These two references indicate that ADIPOQ is present in the adipose tissue secretome in humans and rats.

e. Aggregation

The final aim of Fat&MuscleDB is to aggregate the data. For this, by filling a form, users choose one of the criteria in the list of criteria related to adipose and muscle tissues growth, as well as a species, a breed, a biological sample, or a cell line. Then, users receive the list of found references. They can select or unselect references datasets to aggregate and reverse references having differential abundances data to put the criterion selected on the first condition relative to the other condition. A unique code is given for each aggregation calculated so the user can return on it. We report here the example of the aggregation of data related to the criterion "Proteins secreted by muscles". Currently, this aggregation uses data from 5 references on genomic expression from rat or human. The output list contains 330 UniProt accessions, among them 13 UniProt accessions are found in two references from human, which highlight the added-value of aggregation (Figure 3). Another example of aggregation with the criterion "Muscular hypertrophy from genetic origin" is illustrated in Figure 3. We retrieved 30 samples comparisons given by 10 references from GEO datasets and publications. The aggregation of these references indicates that there are 22,076 proteins/transcripts which are stable in the context of muscular hypertrophy from genetic origin, whereas 1,006 are decreased and 1,021 are increased. An accession may be found as stable, increased or decreased as for example "CO1A2_BOVIN". This indicates that its expression may be dependent of breed, gender, age, tissue or physiological status used in the aggregated references. Indeed, it is found over-expressed in the microarray data from the "GSE5659" GEO dataset through the samples of Piedmontese/Hereford cross relative to Hereford/Wagyu cross at 7 months, whereas it is found stable over these same samples at 3 months and 12 months. This indicates that for

aggregation authors have to carefully select the dataset to be merged. To explore in depth genes or proteins lists obtained through Fat&MuscleDB it is possible to use informatics resources such as ProteINSIDE to identify central or relevant genes or proteins related to a criterion.

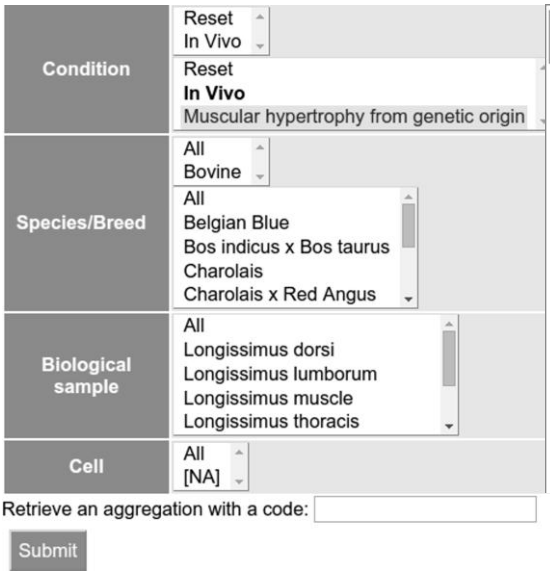
a



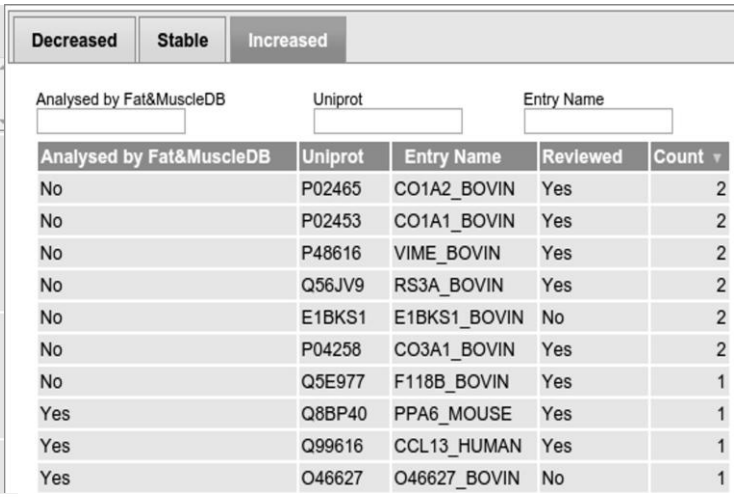
b

Analysed by Fat&MuscleDB	Uniprot	Entry Name	Reviewed	Count
No	P01023	A2MG_HUMAN	Yes	2
No	P51884	LUM_HUMAN	Yes	2
No	P36955	PEDF_HUMAN	Yes	2
No	P01033	TIMP1_HUMAN	Yes	2
No	P09871	C1S_HUMAN	Yes	2
No	P00749	UROK_HUMAN	Yes	2
No	P07711	CATL1_HUMAN	Yes	2
No	P06396	GELS_HUMAN	Yes	2
No	Q12805	FBLN3_HUMAN	Yes	2
No	P09668	CATH_HUMAN	Yes	2
No	P07858	CATB_HUMAN	Yes	2
No	P02788	TRFL_HUMAN	Yes	2
No	P09486	SPRC_HUMAN	Yes	2
No	Q96HF1	SFRP2_HUMAN	Yes	1
No	P25809	KCRU_RAT	Yes	1

c



d



Analysed by Fat&MuscleDB	Uniprot	Entry Name	Reviewed	Count
No	P02465	CO1A2_BOVIN	Yes	2
No	P02453	CO1A1_BOVIN	Yes	2
No	P48616	VIME_BOVIN	Yes	2
No	Q56JV9	RS3A_BOVIN	Yes	2
No	E1BKS1	E1BKS1_BOVIN	No	2
No	P04258	CO3A1_BOVIN	Yes	2
No	Q5E977	F118B_BOVIN	Yes	1
Yes	Q8BP40	PPA6_MOUSE	Yes	1
Yes	Q99616	CCL13_HUMAN	Yes	1
Yes	O46627	O46627_BOVIN	No	1

Figure 3. Aggregation of "omics" data relative to two criteria in Fat&MuscleDB. (a, c) The form allows selecting references to aggregate according to different settings: *in vitro* or *in vivo* experiments, species and breed, biological sample, and cell lines. (a) The criterion selected is "Proteins secreted by muscles". (c) The criterion selected is "Muscular hypertrophy from genetic origin". (b) Extract of the list of proteins which are present in "Proteins secreted by muscles". (d) Extract of the list of proteins whose abundance increases in "Muscular hypertrophy from genetic origin". (b, d) Each line corresponds to a protein with multiple indications: if it is retrieved by us ("Yes") or according to authors' interpretation ("No"), its UniProt accession, its Entry name, if it has been reviewed by the UniProt consortium or not, and the number of times it is retrieved. The duplicate accessions in a reference are only counted once.

f. Conclusion

Fat&MuscleDB is a novel and unique database which brings "omics" research on muscle and adipose tissues into one location and which can be used through the Internet by a web interface. It connects beef industry or human health issues to the new biological discoveries in "omics" research. We will use the

Fat&MuscleDB to identify genes or proteins involved in the lean-to-fat ratio in cattle. As it remains a lot of data to integrate, we continue to feed Fat&MuscleDB. This strategy and this database will undoubtedly reduce avoidable duplication of experimental studies by a better use of available knowledge.

Acknowledgements

This work was supported by the regional council of Auvergne in France through the regional information system Lifegrid. The authors wish to thank Anne de la Foye for her help in statistical analysis.

References

[1] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 41 (Database issue), p. D8–D20, 2013.

[2] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis and A. Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, 41 (D1), p. D991–D995, 2013.

[3] D.R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A.M. Chinnaiyan. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia N. Y. N*, 6 (1), p. 1–6, 2004.

[4] M. Claessens, L. Contor, R. Dhonukshe-Rutten, L.C. De Groot, S.J. Fairweather-Tait, M. Gurinovic, B. Koletzko, B. Van Ommen, M.M. Raats and P. Van't Veer. EURRECA—Principles and Future for Deriving Micronutrient Recommendations. *Crit. Rev. Food Sci. Nutr.*, 53 (10), p. 1135–1146, 2013.

[5] D. Baron, E. Dubois, A. Bihouée, R. Teusan, M. Steenman, P. Jourdon, A. Magot, Y. Péréon, R. Veitia, F. Savagner, G. Ramstein and R. Houlgatte. Meta-analysis of muscle transcriptome data using the MADMuscle database reveals biologically relevant gene patterns. *BMC Genomics*, 12 (1), p. 113, 2011.

[6] C.J. Porter, G.A. Palidwor, R. Sandie, P.M. Krzyzanowski, E.M. Muro, C. Perez-Iratxeta and M.A. Andrade-Navarro. StemBase: a resource for the analysis of stem cell gene expression data. *Methods Mol. Biol. Clifton NJ*, 407, p. 137–148, 2007.

[7] M. Bonnet, I. Cassar-Malek, Y. Chilliard and B. Picard. Ontogenesis of muscle and adipose tissues and their interactions in ruminants and other species. *Anim. Int. J. Anim. Biosci.*, 4 (7), p. 1093–1109, 2010.

[8] T. Romacho, M. Elsen, D. Röhrborn and J. Eckel. Adipose tissue and its role in organ crosstalk. *Acta Physiol.*, 210 (4), p. 733–753, 2014.

[9] T.N. Petersen, S. Brunak, G. von Heijne and H. Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, 8 (10), p. 785–786, 2011.

[10] S. Davis and P.S. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23 (14), p. 1846–1847, 2007.

[11] UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 42 (11), p. 7486, 2014.

[12] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11 (10), p. R106, 2010.

[13] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer and A. Conesa. Differential expression in RNA-seq: A matter of depth. *Genome Res.*, p. gr.124321.111, 2011.

[14] J.S. Hamid, P. Hu, N.M. Roslin, V. Ling, C.M.T. Greenwood and J. Beyene. Data Integration in Genetics and Genomics: Methods and Challenges. *Hum. Genomics Proteomics HGP*, 2009.

[15] I. Cassar-Malek, F. Passelaigue, C. Bernard, J. Léger and J.-F. Hocquette. Target genes of myostatin loss-of-function in muscles of late bovine fetuses. *BMC Genomics*, 8, p. 63, 2007.