Amazonia!: An Online Resource to Google and Visualize Public Human whole Genome Expression Data

Tanguy Le Carrour[1,2], Said Assou[1,2], Sylvie Tondeur[1,2,3], Ludovic Lhermitte[4], Ned Lamb[5], Thierry Rème[1,2], Véronique Pantesco[1,2], Samir Hamamah[1,2,6,7], Bernard Klein[1,2,6] and John De Vos*[,1,2,6,8]

[1]CHU Montpellier, Institute for Research in Biotherapy, Hôpital Saint-Eloi, Montpellier, F-34000 France; [2]INSERM, U847, Montpellier, F-34000 France; [3]CHU Montpellier, Laboratoire d'Hématologie, Hôpital St-Eloi, F-34000 France; [4]CNRS UMR 8147 and Department of Haematology, Hôpital Necker-Enfants-Malades Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris-5 Descartes Paris, F-75015 Paris; [5]Institute of Human Genetics, CNRS UPR-1142, Montpellier, F 34000 France; [6]Université MONTPELLIER1, UFR de Médecine, Montpellier, F-34000 France; [7]CHU Montpellier, Département de Médecine et Biologie de la Reproduction, Hôpital Arnaud de Villeneuve, Montpellier, F-34000; [8]CHU Montpellier, Unité de Thérapie Cellulaire et Génique, Hôpital Saint-Eloi, Montpellier, F-34000 France

Abstract: Available transcriptome data accumulate in public repositories, individual web pages or as various supplemental data, but these published data cannot be routinely accessed. We have developed the web based tool Amazonia! to overcome this hurdle and provide the possibility to query and to visualize the expression of a given gene in representative and selected human transcriptome datasets. This expression atlas provides expression bar plots for single genes, across samples selected from a wide range of normal tissues and malignancies, including pluripotent stem cells. When produced by the same platform type, datasets were renormalized and combined in order to generate series of several hundreds samples. Samples types are colored and ordered, and grouped in thematic pages for ease of navigation. We also integrated gene lists provided by original publications describing these microarray data, allowing the scientific community to challenge the expression of genes in datasets other than those for which they were initially published. To illustrate the powerfulness of this simple tool, we show how Amazonia! reveals the specific expression of the tight junction protein Claudin 6 in human embryonic stem cells and human induced stem cells (iPS), or the tissue specific expression of some chemokines and their receptors such as CCL16 in liver and CX3CR1 in central nervous system samples. Thus, Amazonia! advantageously complements large public repositories by providing a simple way to query a compilation of selected human transcriptome data.

The tool is freely available at http://www.amazonia.transcriptome.eu/

Keywords: Pluripotent stem cells, Embryonic stem cells, Therapeutics, Cell reprogramming, Cell proliferation, Cell differentiation.

INTRODUCTION

Microarray technology is a major technical breakthrough that can monitor the expression of a whole genome within a single experiment. Information provided by this new technology is frequently shared with the scientific community and is freely accessible. But paradoxically, and this is a direct consequence of the massive quantity of data generated, most researchers do not access to these data on a routine basis. Indeed, many aspects of microarray data mining require substantial know-how and labour time, such as unsupervised or supervised analyses. But microarrays results must also be viewed as a massive expression repository, providing hundred of thousands of virtual northern blots. Several web sites already propose to access these expression profiles on a gene per gene basis. Gene Expression Omnibus (GEO) [1] is a public repository of microarray data, and provides an expression bar plot for each probe in all datasets present in the database. However, the heterogeneity in the datasets, covering all biological fields, from yeast to human, from normal to malignant cells, from primary tissues to *in vitro* stimulated cell lines, discourages users from investigating the expression of a given gene in this transcriptome repository. For instance, the keyword "embryonic" yields 1161 different GEO data-sets and 3 145 499 GEO profiles. Other web sites have focused on datamining such as the cancer-centered Oncomine [2]. Finally, many sites provide only access to expression profiles obtained in their lab as supplemental data in connection with a publication, such as the GNF SymAtlas [3] or the Stanford Microarray Database [4].

Another concern is that most transcriptome analyses published provide lists of genes differentially expressed between different sample groups. Like the raw data, these lists are infrequently consulted because they are provided as printed mater or supplemental data. This is unfortunate since these lists of genes reflect the conclusions of the authors that carried out the analyses, and should be challenged and confronted with alternative interpretations.

*Address correspondence to these authors at the Institute for Research in Biotherapy, Hôpital Saint-Eloi, 80 Avenue Augustin Fliche, 34295 Montpellier Cedex 5, France; Tel: +33 (0)4 67 33 01 91; Fax: 33 (0)4 67 33 79 05; E-mail: john.devos@inserm.fr

1875-0362/10

Therefore, we developed Amazonia!, a web-based atlas of human gene expression that compiles a selection of publicly available transcriptome datasets, that is freely accessible through a user friendly interface to the research community. The database is gene centered, organized in thematic pages and provides enhanced sample identification features such as color and rank attributes. A list manager can be used to scrutinize the expression of genes from a published list across other datasets.

METHODS AND IMPLEMENTATION

Amazonia!, subtitled "Explore the jungle of microarrays results", was developed with open source software : data are stored in a MYSQL database, the web pages were created using PHP and are served by an Apache web server. Its aim is to provide a straightforward way to retrieve and visualize as bar plots the expression of a gene across large series of samples (see outline of the Amazonia! database in Fig. **1**). These series were carefully selected and organized into thematic pages so that users can focus on their sample of interest (normal tissues or a given cancer for instance). Amazonia! is freely accessible at http://www.amazonia.transcriptome.eu.

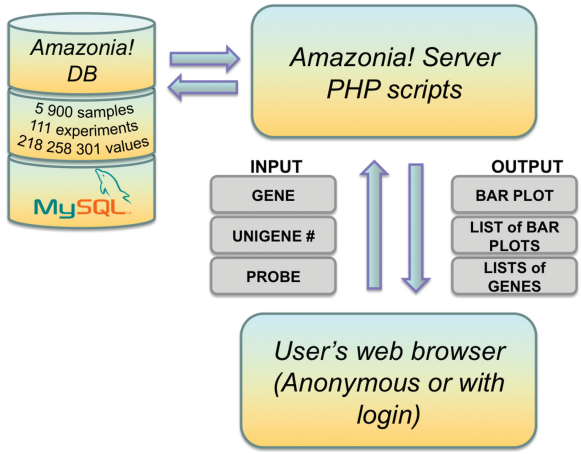

Fig. (1). Outline of the Amazonia! database and web server.

We selected relevant datasets that were made publicly available, either in public microarray repositories (GEO and ArrayExpress)[1, 5] or as supplemental data in various publications. For microarray results, the y-axis of the bar plot is the normalized microarray signal value. This is an arbitrary unit, proportional to the mRNA quantity in the initial preparation. This value depends on the microarray format and the normalization used. It is comparable between series of samples only when the same microarray and the same normalization scheme have been used. Transcriptome data were obtained with various technologies, including massively parallel signature sequencing (MPSS), cDNA spotted microarrays and Affymetrix microarrays, the latter being the most frequent source of data in our collection. In the near future, deep sequencing digital gene expression (DGE) data will be added. In this case, the y-axis would be the tag per million count. When raw Affymetrix files were accessible, the data were normalized and summarized with the MAS5 algorithm starting from the raw files, using the Expression Console software (Affymetrix, Santa Clara, CA) by global scaling with a trimmed mean target intensity value (TGT) for each array arbitrarily set to 100. In this way, Affymetrix data obtained from different platforms in different places are directly comparable. For each expression bar plot, the provenance of each sample is documented on the web page with a direct link to the public repository web page, or alternatively to the publication abstract in Pubmed, and therefore provides a mean to access all experimental details on the original source web page.

RESULTS AND DISCUSSION

Bar Plots

This expression atlas provides expression bar plots for single genes, across different collections of samples. Genes are accessed either by keywords (including gene name aliases), Gene identification (ID), UniGene ID or probe ID (Fig. **2A**). If the expression of one gene is queried by several probes, every probe will be displayed. For keywords, a disambiguation page lists all genes containing this keyword either within the abbreviation, alias or full name (Fig. **2B**). Selection of a gene in this list will call the expression bar plots (Fig. **2C**). The bar plot page contains the abbreviation of the gene, name, aliase(s), chromosomal location, gene ID, links to external gene annotation resources (Entrez Gene, iHOP, PubMed, OMIM and Gene Cards) and the bar plot(s). By default, the expression of a gene will be profiled in a series of normal samples from embryonic, fetal and adult tissues ("Human body index"). This series is a virtual northern blot covering more than 230 human samples. These samples are obtained from more than ten (13 at the time of writing) different studies normalized by the same method, thereby allowing direct signal values comparisons. Each bar plot is a png image incorporating some key parameters such as the gene abbreviation, probeset ID, microarray type, hybridization protocol and normalization method. By a simple click, a higher-solution image can be called out and shown in a new window.

An essential aspect of Amazonia! is the careful sample color and rank attribute so that (i) samples that have a similar histological origin have a similar color, (ii) samples are grouped according to their primitive embryonic layer origin and (iii) normal and malignant samples for a same tissue are differentiated by a shade variation. When the microarray data has been obtained on an Affymetrix GeneChip system, the "Detection call" information provided by Affymetrix software, which indicates whether a transcript is reliably detected (Present), ambiguous (Marginal) or not detected (Absent), is included in the color scheme, so that only the samples with Detection calls Present or Marginal are colored whereas the samples with a Detection call Absent are white. It is notorious that some probesets do not provide signals above background: using the color scheme, these probesets are immediately evident since most bars are white. Signals and, if relevant (i.e. for Affymetrix microarray data), Detection p-values that were used to generate the bar plot can be downloaded by clicking on the "view data" icon (Fig. **2C**). With this color and ordering strategy, one can have a comprehensive view of the sample collection in a bar plot at a glance. For instance, the gene Ribosomal protein L13 (RPL13) is expressed at high level in all tissues from the

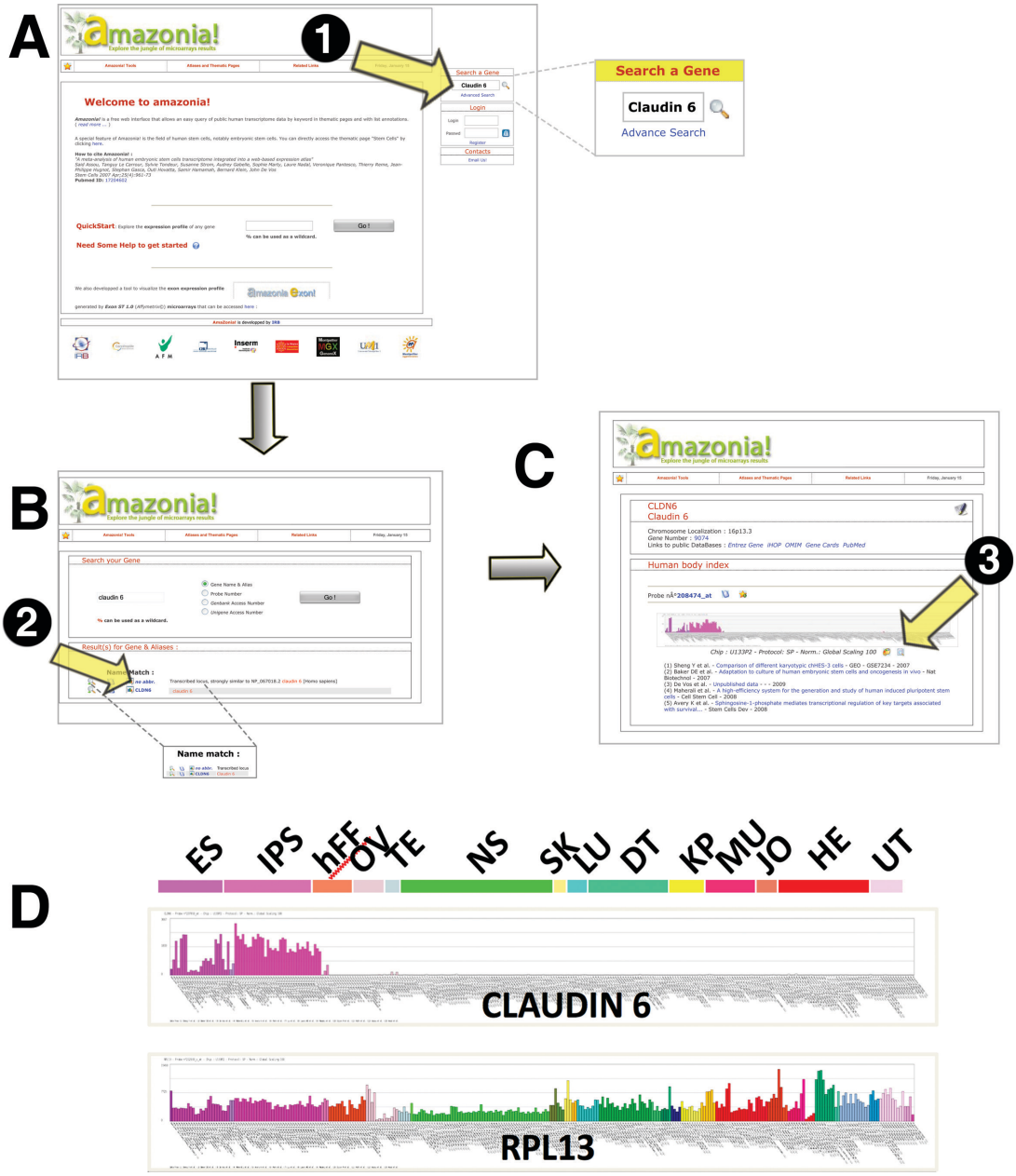

Fig. (2). Amazonia! main input and disambiguation of keywords. (**A**) To obtain an expression profile for a given gene, a keyword (generic character % can be used), or a probeset number or a UniGene ID or a Gene ID can be used as input (1). (**B**) If a keyword is used and matches different genes, a first page is displayed that lists all the genes that contain the keyword in its official abbreviation, alias or full name. An exhaustive list can be obtained by clicking on the "…" link at the bottom of the web page. The user must then click on the selected gene to obtain the corresponding expression bar plot. (**C**) By clicking on the selected gene (2), all bar plots corresponding to the selected gene are displayed. The data corresponding to the bar plot can be downloaded by clicking on the "view data" icon (3). (**D**) The bar plots of CLAUDIN 6 and RPL13 are shown as examples of one highly specific gene, highly expressed in pluripotent stem cells (human embryonic stem cells and pluripotent stem cells) and one housekeeping gene, broadly expressed across all tissues, respectively. ES: human embryonic stem cells; IPS: induced pluripotent stem cells; hFF: human foreskin fibroblasts; OV: ovary & oocytes samples; TE: testis; NS : nervous system; SK: skin; LU: normal lung; DT: digestive tract; KP: kidney & prostate; HM: heart & muscle; JO : joint; HE: normal hematological samples; UT: uterus.

"Human body index" series, with the only exception being a low expression in oocytes (Fig. **2D**). Hence, RPL13 appears as a ubiquitously expressed gene, a common feature for ribosomal proteins. Nine different probesets (PS) interrogate this gene, but two display bar plots with a majority of white bars, i.e. an "absent" Detection call (PS 1565758_at and PS 1565759_at), which could be explained by a low signal/noise ratio. By contrast, the gene Claudin 6 (CLDN6) coding for a member of tight junction proteins is highly expressed in human embryonic stem cells and human induced pluripotent

stem cells (iPS) but completely absent from the other adult or fetal tissues from the "Human body index" series (Fig. **2D**). A glimpse at this bar plot is sufficient for concluding that CLDN6 is a bona fide pluripotency gene.

Thematic Pages

Whereas the goal of public repositories such as GEO or Array Express is comprehensiveness, the goal of Amazonia! is instead to select a limited panel of microarray experiments that are representative of the expression of our genome in most normal tissues and some of the most frequent malignancies. As of October 1st, 2009, more than 218 258 301 individual expression values have been introduced into the database, from more than 5 900 samples in 111 different experiments. These experiments have been ordered into several thematic pages: "General", "Hematology", "Cancer", "Reproductive Tract" and "Stem Cells". These thematic pages can be accessed through a cascading menu, each menu having a sub menu (Fig. **3A**). For example, the Cancer menu is divided at present in the following sub menus: "Kidney and prostate cancer", "Lung cancer", "Glioma", "Breast cancer", "Colon cancer" and "Hepatocarcinoma". The range of cancer types will be extended in the future. By switching the thematic page, the queried gene is kept, but the sample collections vary. This is illustrated for the genes ACTG1, PBX1 and MAGEA6. The Actin gamma 1 gene (ACTG1) is ubiquitously expressed either in a normal embryonic, fetal and adult tissues series, an acute lymphoid leukemia (ALL) samples series or a lung carcinoma samples series as shown by the three bar plots corresponding to the same gene by only switching the thematic page (Fig. **3B**, **C** and **D** respectively). By contrast, the Pre-B-cell leukemia homeobox 1 gene (PBX1) is preferentially expressed in human embryonic stem cells, but also in a subset of ALL that carries a translocation involving PBX1 and resulting in its specific overexpression. Finally, the Melanoma antigen family A, 6 gene (MAGEA6) is expressed in testis, ALL cell lines and some lung cancer samples, highlighting some essential features of this cancer testis gene. Similarly, the expression of any gene can be queried in various series of samples, by simply changing the thematic page in the drop down menu.

Lists of Genes

A list manager offers the possibility to obtain the expression bar plots for tens or hundreds of genes out of a list. Two kinds of gene lists are proposed: (i) a group of genes from a

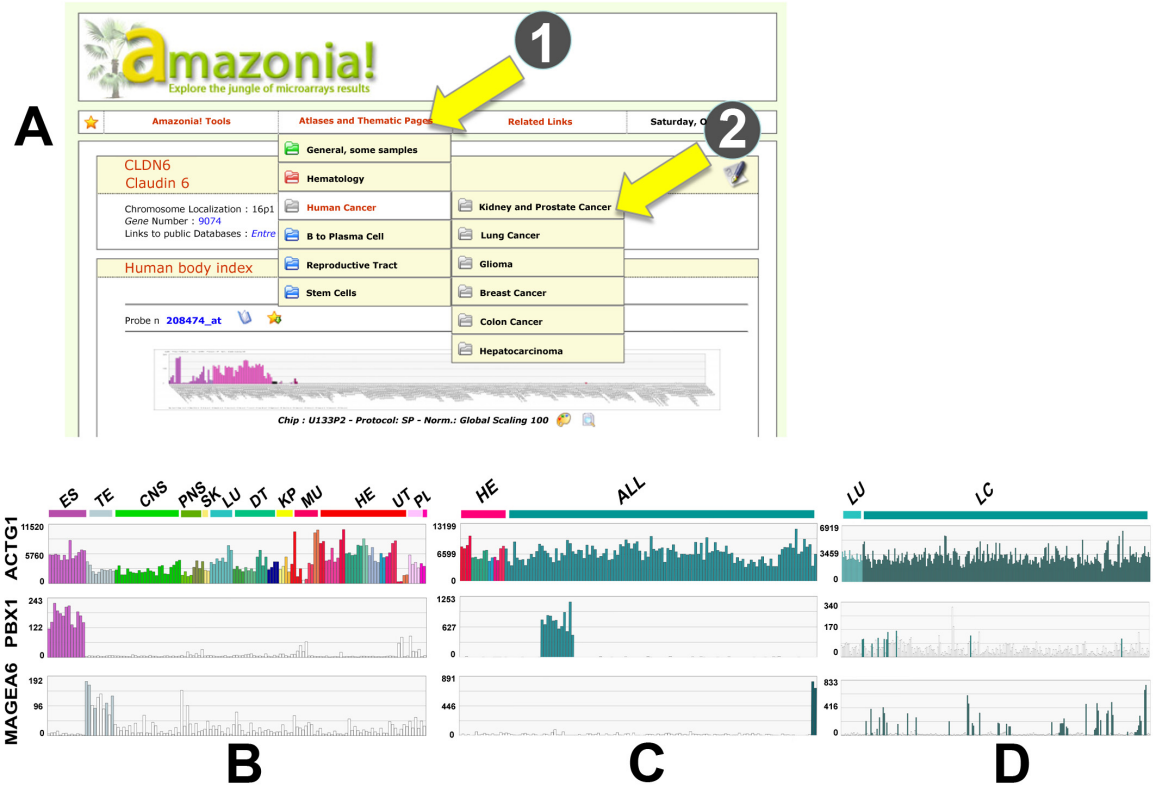

Fig. (3). Thematic pages. (**A**) Thematic pages can be accessed through a cascading menu, each menu having a sub menu. For the sake of illustration, three expression profiles obtained with Amazonia! for an ubiquitous gene, Actin G1 (ACTG1), a transcription factor associated with some subtypes of malignancies, pre-B-cell leukemia transcription factor 1 (PBX1) and a cancer/testis antigen, melanoma antigen family A, 6 (MAGEA6), are illustrated in three samples series: embryonic and adult normal tissues (**B**) (U133A microarray data), acute lymphoblastic leukemia (ALL) (**C**) (U133A microarray data) and lung cancer (LC) (**D**) (U95Av2 microarray data). One can observe that the PBX1 transcription factor is typically overexpressed in a subset of ALL samples through a reciprocal translocation, but is also highly expressed in human embryonic stem cells. MAGEA6 is detected in most normal testis samples and is highly expressed in one T-ALL cell line and in a subset of lung cancer samples. ES: human embryonic stem cells; TE: testis; CNS : central nervous system; PNS : peripheral nervous system; SK: skin; LU : normal lung; DT: digestive tract; TH : thyroid; KP: kidney & prostate; HM: heart & muscle; HE: normal hematological samples; UT: uterus; PL: placenta.

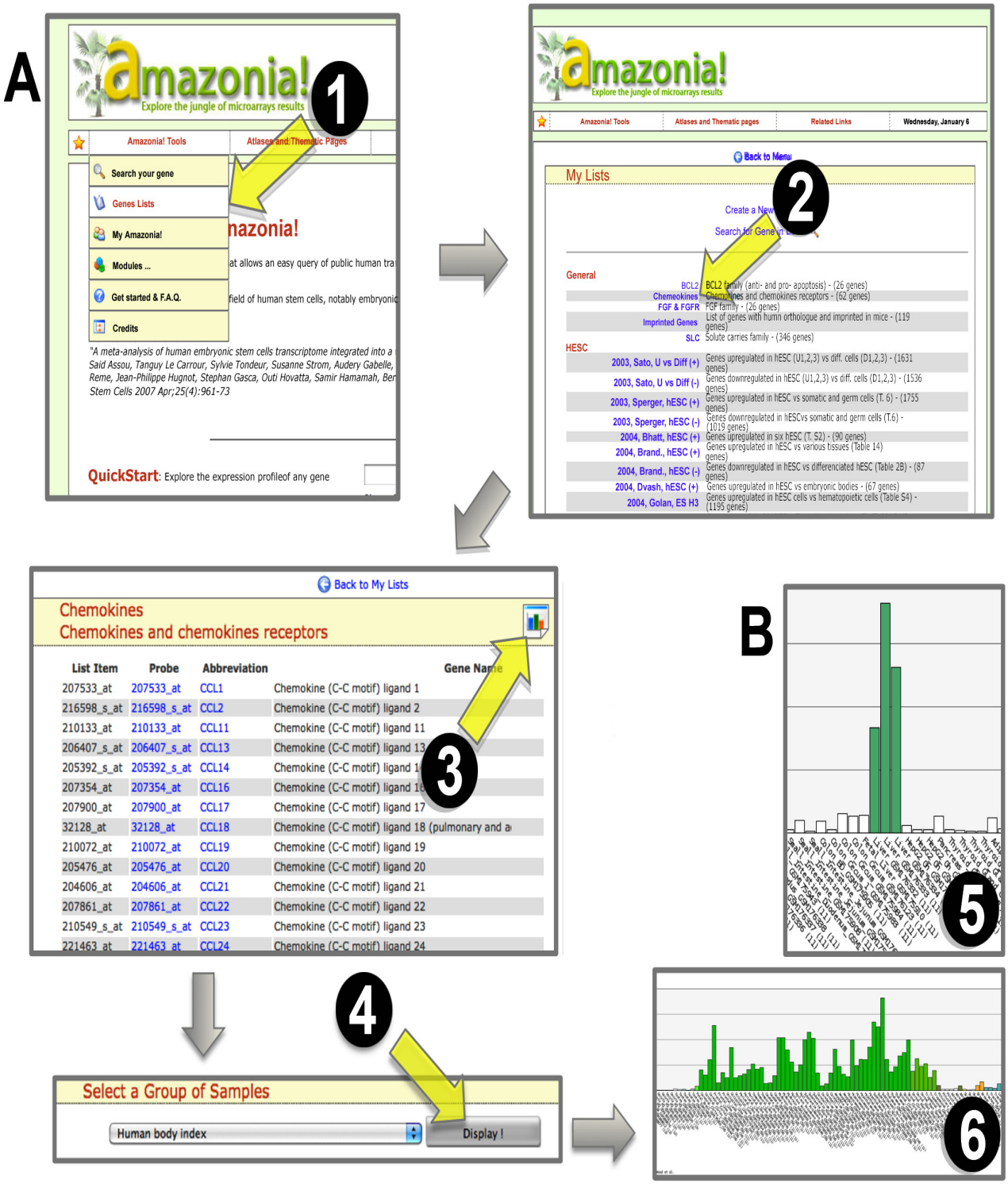

Fig. (4). Generation of a list of bar plots. (**A**) Several published lists of genes can be accessed *via* the gene list module found in the Amazonia! Tools menu (1). By clicking on a given list (2), all genes are displayed in the order provided by the author of the list, usually by decreasing fold change or alphabetical order. By clicking on the bar plot tool (3), and selecting any of the samples series available (by default the Human Body Index series), the expression bar plots for each gene or probeset of the list is displayed (4). (**B**) A partial view of the bar plots of the "chemokine" list, with a highlight on the exquisite expression of the chemokine CCL16 in liver (5) and the chemokine receptor CX3CR1 in central nervous system samples (6).

gene family or from a biological pathway, lists that are included in the "General" section, or (ii) lists of genes found differentially expressed between two cell types and previously published, organized in sections corresponding to the Amazonia! thematic pages. But most interestingly, every registered user (see below, "Registering") can create its own list by following the menu "Amazonia! tools/Lists of genes/Create a new list" and by pasting genes (gene ID), UniGene (UniGene ID) or probesets in the dedicated field. Lists of genes can then be displayed as bar plots. For instance, the "Chemokine" list contains 62 genes that include chemokines and their receptors (Fig. **4**). By selecting the bar plot icon, and then for example the "Human body index" series of samples (the default series of samples), all bar plots corresponding to all probesets from all 62 genes will be displayed on two web pages (the display is limited to 50 bar plots per page). This is a powerful tool to uncover specific expression patterns by simple visual scanning. We note for example that some chemokines have a very selective expression, such as the high expression of CCL16 in liver, CX3CR1 in central nervous system samples, CCL17 in activated B lymphocytes and CCL27 in skin samples (Fig. **4B** and data not shown, but freely accessible on the web site). Hence, challenging a published list of genes in a new dataset, independent of the one that was used in the original publication, is becoming an easy task.

Registering and Private Access

Amazonia! is freely accessible and registering is not mandatory. It is however strongly recommended as this opens the way for users to gain access to several Amazonia! tools such as the "create a new list" (see above) or another tool to generate a user's own series of samples, by combining samples from various experiments that have been done on the same experimental platform, microarray format and normalization (typically within Affymetrix datasets). This is a convenient way to build figures to illustrate a study without the need of importing the data into a graphing software [6, 7]. A recent development is the inclusion of large collections of human pluripotent stem cells, both human embryonic stem cells and human induced pluripotent stem cells, as well as adult stem cell samples such as CD34 positive hematopoietic stem cells [8, 9].

CONCLUSION

Thus, Amazonia! fills a gap in the on-line microarray tools by providing a simple but powerful way to query a collection of representative microarray data covering a wide range of normal and malignant human samples.

ACKNOWLEDGEMENTS

We thank Cyril Berthenet and Gilles Palenzuela for their help in the early steps of this project and Qiang Bai for its careful reading of the manuscript. This work was supported by the Association Française contre les Myopathies (AFM) and the Cancéropole Grand Sud-Ouest.

REFERENCES

[1] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, pp. 207-210, 2002.

[2] 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*, vol. 6, pp. 1-6, 2004.

[3] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 6062-6067, 2004.

[4] J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock, and C. A. Ball, "The stanford microarray database: implementation of new analysis tools and open source release of software," *Nucleic Acids Res.*, vol. 35, pp. D766-D770, 2007.

[5] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S. A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma, "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Res.*, vol. 37, pp. D868-D872, 2009.

[6] J. De Vos, D. Hose, T. Reme, K. Tarte, J. Moreaux, K. Mahtouk, M. Jourdan, H. Goldschmidt, J. F. Rossi, F. W. Cremer, and B. Klein, "Microarray-based understanding of normal and malignant plasma cells," *Immunol. Rev.*, vol. 210, pp. 86-104, 2006.

[7] M. Jourdan, A. Caraux, J. De Vos, G. Fiol, M. Larroque, C. Cognot, C. Bret, C. Duperray, D. Hose, and B. Klein, "An *in vitro* model of differentiation of memory B cells into plasmablasts and plasma cells including detailed phenotypic and molecular characterization," *Blood*, vol. 114, pp. 5173-5181, 2009.

[8] S. Assou, T. Lecarrour, S. Tondeur, S. Ström, A. Gabelle, S. Marty, L. Nadal, V. Pantesco, T. Reme, J. P. Hugnot, S. Gasca, O. Hovata, S. Hamamah, B. klein, and J. De Vos, "A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas," *Stem Cells*, vol. 25, pp. 961-973, 2007.

[9] S. Assou, D. Cerecedo, S. Tondeur, V. Pantesco, O. Hovatta, B. Klein, S. Hamamah, and J. De Vos, "A gene expression signature shared by human mature oocytes and embryonic stem cells," *BMC Genomics*, vol. 10, pp. 10-24, 2009.

Received: November 11, 2009 Revised: December 23, 2009 Accepted: December 23, 2009